ISBN: 978-93-6773-974-7

DETECTING CYBERBULLYING: A COMPREHENSIVE SURVEY

MANISH JOSHI
DEPARTMENT OF INFORMATION TECHNOLOGY
BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY
LUCKNOW, INDIA
edu.manishjoshi@gmail.com

DR.DHIRENDRA PANDEY
DEPARTMENT OF INFORMATION TECHNOLOGY
BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY
LUCKNOW, INDIA
prof.dhiren@gmail.com

KEYWORDS

CYBERBULLYING, DEEP LEARNING, MACHINE LEARNING, SOCIAL MEDIA

ABSTRACT

The usage of social media over the past decade has made sharing of one's thoughts and views anonymously without being judged but others. However, this anonymity is also used by many to harm other individuals. People who hide behind the mask of anonymity, hiding their identity can reach their victim at any time and place. This has led to cyberbullying being a widespread problem, especially on social media websites and seeing the negative effect of cyberbullying there is an increase demand for strategies to tackle cyberbullying. This study aims to provide a detailed analysis of various detection methods for cyberbullying and also analyze the research trends, gaps and future prospects. This study consists of 61 research studies that focuses on cyberbullying detection using various methods.

1. INTRODUCTION

The internet has had an important impact on mankind. It has changed how people interact and socialize with each other and it has been on rise over the past decade. The medium for this kind of human interaction over the internet is termed as Social Media. Social Media website provide a platform for interaction and communication between individuals, it scales geographical bound and culture to connect individual, communities with similar or dissimilar interests, it also help individuals show their creativity. People can share knowledge and information on various topics and attract discussion on them. It is also used as tools for learning new skills, it helps bringing educators and professional with students and learners. Though with these seeming benefits there also lies a dark side to it. Cyberbullying is one such problem that has been plaguing social media. According to UNICIEF Cyberbullying is defined as "bullying with the use of digital technologies. It can take place on social media, messaging platforms, gaming platforms and mobile phones. It is repeated behaviour, aimed at scaring, angering or shaming those who are targeted" [1], American Psychological Council Association and White House has identified cyberbullying as a serious health concern. National Crime Prevention Council, has estimated that more than 40% of the US teens have been bullied on social media websites [2]. Cyberbullying victim often show symptoms of lower self-esteem, anxiety, depression and loneliness. Victims of cyberbullying often show mental health problems and in some cases they might attempt suicide.

The main difference in traditional bullying and cyberbullying is that cyberbullying is not restricted by geographical boundaries and can occur at any time. The traditional ways to combat cyberbullying such as standard and guidelines, using profane list, doesn't do well on social media, they are often forced to resort to manual effort to identify the cyberbullying instances which often time and labour consuming, and very difficult to scale. This necessitates the framework to detect cyberbullying instances on various social media. Cyberbullying can happen at any time, and reach their victims even when he or she is alone. Second, due to anonymous nature of several social media bullies often create fake profile and post message and images to target their victim and reach a wide range of audience, it is very difficult and challenging to identify these cyberbullies. Lastly, it is very difficult to correctly appraise a message or post as cyberbullying as they can be ambiguous and model can identify it as cyberbullying when it's not and vice versa. This study examines the classification of internet bullying content using different approaches. Key contribution of the studies is mentioned below.

- We did a comprehensive review of the existing studies. We selected 61 studies which focused on detection of cyberbullying using various techniques.
- We identified various models and techniques which are used in cyberbullying detection. Our study also provided a detailed summary of each of the study, which had information like classifiers used, best classifier in study etc.
- We provided an in-depth analysis on various dataset used in detection of cyberbullying. This analysis helps to shed the light in availability, diversity and quality of dataset which are being used for evaluating cyberbullying detection models. It also highlights the need for standardized, diverse and well labelled dataset to ensure we get reliable results.
- Finally, by identifying the techniques/models, and datasets utilized in these studies, we offer valuable guidelines for researchers and practitioners working on cyberbullying detection using machine learning and deep learning techniques future cyberbullying detection research.

The study is structured as follows: Section 2 details on the methodology used for review by stating the research topics identified for this study review. Finally, section 3 includes results and discussion, followed by section 4 conclusion, which identifies prospective research gaps and future guidelines.

2. METHODOLOGY

This section describes the methods employed for the literature search in this study. We wish to answer the following research problems.

- Rp 1: Which techniques are employed most commonly for detection of cyberbullying on social media?
- Rp 2: What is the source, type and domain of data that is being utilized in cyberbullying detection? We also analysed whether the agreement rate between labellers was taken into account or not while labelling of data and, whether labellers had any expertise in the field of cyberbullying or not.
- Rp 3: What guidelines for future cyberbullying detection research may be derived from the reported findings of this review?

The retrieval of published articles on cyberbullying prediction models involved two parts. The initial phase consisted of a search for credible academic resources and search engines. Relevant publications were retrieved using the following search engines and academic databases: ACM Digital Library, Clarivate Analytics' Web of Science, Scopus, IEEE Xplore, Science Direct and Springer Link. Cyberbullying,

cyber bullying, aggressive behaviour, cyberbullying detection and cyberbullying models were developed as the primary keywords for the literature search. The search-retrieved articles were evaluated to determine whether they fit the inclusion criteria. Based on the inclusion criteria, for an article to be selected for the survey, it should describe an empirical study that predicts cyberbullying on Social Media sites. If not, the study would be excluded from consideration. Numerous articles were rejected due to their titles. The abstract and conclusion portions of the papers were reviewed to confirm that they met the screening criteria, and those that did not were omitted from the survey. This study's inclusion criteria were research studies published in the last 13 years, research demonstrating cyberbullying monitoring and detection on social media platforms, and research focusing on the implementation of supervised and unsupervised machine learning techniques. Research studies demonstrating the implantation of deep learning techniques to the identification of cyberbullying on social media platforms and research utilising hybrid models and techniques to detect cyberbullying on social media platforms. Exclusion criteria for this study was studies that lack appropriate empirical analysis or comparisons to relevant standards and Cyberbullying detection studies that are reviews, surveys, or theoretical concepts with no implementations. Table II provides the detailed summary of the studies included.

3. RESULTS AND DISCUSSION

3.1 RQ1: WHICH TECHNIQUES ARE EMPLOYED MOST COMMONLY FOR DETECTION OF CYBERBULLYING ON SOCIAL MEDIA?

The research shows that a variety of approaches have been used to identify and anticipate online cyberbullying. There were few reports of early attempts to use supervised machine learning for Cyberbullying Detection on text information by [5] and [6]. Following that, supervised machine learning techniques such as Support Vector Machine, Naive Bayes, Ridge Regression, Decision Tree, Random Forest, Stochastic Gradient Descent, Bayes Algorithm, Extra Tree, Logistic LASSO Regression, Essential Dimensions of Latent Semantic Indexing, and Gradient Boosting were incorporated on a range of social media platforms such as Facebook, Formspring.me, Myspace, Kaggle, Slashdot, Kongregate, X (formally known as Twitter), Wikipedia and Reddit. Traditional Cyberbullying Detection techniques (vocabulary, rule-based techniques) yield inferior results when compared to the presented methods. Similarly, increasing use of techniques such as Fuzzy Logic, Genetic Algorithm, Maximum Entropy, Neural Network, and others, as well as

Ensemble methods such as Adaboost, Boosting, and Bagging, has been demonstrated for Cyberbullying Detection on social media for all types of multimedia, including photos, video content, audio content, and texts. Data for the implementation was gathered using a variety of social media platforms, including Instagram and YouTube. A few papers also described how to detect cyberbullying using distance-based algorithms like k-Nearest Neighbour and regression-based methods like Logistic Regression, Linear Regression, and Multiple Regression. Unsupervised and semi-supervised approaches for detection of cyberbullying were also investigated in some of the studies. Probabilistic Latent Semantic Analysis, C means clustering, MinHashes, Conditional Random Fields, MinHashes, and Expectation-Maximization, were among some of the techniques used. Many research studies also used hybrid techniques, such as Fuzzy Decision Tree (FDT), Fuzzy C Mean. (FCM) and Fuzzy cyberbullying detection on social media, to improve the system's accuracy, etc[10],[11],[14]. Recently, studies on cyberbullying detection using deep learning techniques such as Long Short-Term Memory (LSTM), marginalized Stacked Denoising Autoencoders (mSDA), Pronunciation-Based Convolution Neural Network, and Convolution Neural Network (CNN), and Bidirectional Long Short-Term Memory (BLSTM), Semantic-Enhanced Marginalized Denoising Auto-Encoders, have been used on messages collected from various social media plat forms to detect cyberbullying.

S.No	Technique	Papers
1	Support Vector machine	35
2	Neural Network	16
3	Naïve Bayes	24
4	Random Forest	14
5	Logistic regression	19
6	Decision Tree	13
7	k-nearest Neighbour	6
8	Fuzzy Logic	4
9	Latent Dirichlet Allocation (LDA)	4
10	Maximum Entropy	1
11	Adaboost	2
12	Latent Semantic Analysis	3
13	Hierarchical agglomerative cluzstering	1
14	Genetic Algorithm	1

TABLE 1.1 CYBERBULLYING DETECTION TECHNIQUES

According to Table I, Support Vector Machine (SVM) is the most commonly used technique. This likely reflects SVM's robustness in handling high-dimensional data and its effectiveness in binary and multi-class classification tasks, However, SVM's reliance on a linear decision boundary can limit its ability to capture complex linguistic patterns in cyberbullying texts, and its high computational cost can present scalability challenges in larger datasets. Naïve Bayes and Logistic Regression are also frequently employed, suggesting a reliance on straightforward, interpretable probabilistic models. These models particularly attractive for its simplicity and efficiency, however they may not capture the underlying complexities of cyberbullying language, which often involves subtle variations in word choice and context.

Neural Networks represent a trend towards more advanced, deep learning approaches, reflecting the field's gradual adoption of models capable of capturing non-linear relationships in language data. While neural networks can be highly effective in detecting nuanced language patterns, their relatively lower count compared to SVM or Naïve Bayes may be attributed to the increased computational resources and expertise. We are seeing trends toward the use of deep learning models in cyberbullying detection, but the size and quality of data still remain a hindrance to utilizing the full capability of these models

1.	[7]	2018	LSTM, SVM, CNN, Random Forest, Naïve Bayes, BLSTM, Logistic Regression	5	F-score, Precision, Recall	Using models based on deep learning and transfer learning, this study examined the detection of cyberbullying across a variety of SMPs and themes.	CNN BLSTM
2.	[8]	2014	MCES (Expert System), Decision Tree, SVM, Naïve Bayes,	10	Area Under Curve	Naive Bayes performed the best when compared to other models	MCES+ Naïve Bayes
3.	[9]	2015	Bayes Point Machine (BPM), KMC, SVM, Decision Tree	10	Sn, True Negative, False Positive, True Positive, False Negative, Specificity	A model for examining different kinds of characteristics for automatically detecting abuse has been proposed.	–
4.	[10]	2014	SVM, kNN, Fuzzy Logic, Max Entropy, Naïve Bayes,	10	Precision, Accuracy, False Positive, False Negative,	Developed a 'Grooming Attack Recognition System' for the real-time detection, evaluation, and regulation of cyber grooming attacks.	Naïve Bayes

5.	[11]	2015	FuzGen, Naïve Bayes, Fuzzy Logic	–	F Score, Accuracy, Precision, Recall	Used Genetic algorithm for optimizing parameters and fuzzy logic to get relevant information form classification of input.	–
6.	[12]	2013	BoW, EDLS, tf-idf	2	True Positive, Precision, Recall	Model for cyberbullying detection which provided better results.	–
7.	[13]	2016	Bayesian hierarchical clustering, SVM, HAC,	–	Recall	Used clustering approach for identifying patterns in cyberbullying involving sexual content.	SAX with sliding window
8.	[14]	2014	Naïve Bayes, mSGD, Random Forest, Logistic Regression, FuzzySVM, Kernel-based FuzzyC-Means	10	F Score, Precision, Recall	Suggested a semi-supervised method which showed better results than traditional models for cyberbullying detection	Kernel based SVM
9.	[15]	2015	Snowball sampling method, Naïve Bayes, AdaBoost, Decision	10	Recall, Accuracy, Precision	Adaboost showed highest accuracy among all.	adaBoost

			Tree, Random Forest				
10.	[16]	2017	Naïve Bayes, kNN, Decision Tree, SVM	–	F Score, Precision, Recall	Presented comparison between direct and indirect bullying, and showed more user are affected by direct bulling	kNN
11.	[17]	2106	LDA, Adaboost, Decision Tree, Random Forest, Extra tree classifier, SVM (variation of SVM), kNN, Naïve Bayes, Multi-Layer Perceptron, Logistic Regression	10	Recall, Accuracy, Precision	Random forest showed the best Accuracy, Precision and Recall	Random Forest
12.	[18]	2017	Random Forest	10	Recall, Precision, Cohen's Kappa, ROC	The proposed approach showed high accuracy for identifying post aggressive and bully content	Random Forest
13.	[19]	2017	SVM,Fuzzy Logic	–	Accuracy	Proposed a model to determine severity of bullying.	–

14.	[20]	2017	Latent Dirichlet Allocation topic modelling, Naïve Bayes	–	Precision, F Score, Recall, Cohen's Kappa	The proposed method improved the classifier's performance for accurate automatic detection.	Naïve Bayes
15.	[21]	2017	Participant-Vocabulary Consistency (PVC) using Alternating Least Squares, snowball sampling	3	Precision	Proposed a weakly supervised Participant-Vocabulary Consistency model.	–
16.	[22]	2017	Naïve Bayes, SVM	–	False Positive, Recall, F score, Precision, True Positive	Used Support Vector Machine and naive Bayes as solution to detect and stop cyberbullying	Support vector machine
17.	[23]	2017	Bayes expectation maximization, Decision Tree, Multivariate Logistic Regression, Maximum Entropy, Winnow2, Bag of Words	10	F Score, Recall, Precision, Accuracy	Implemented logistic regression with word unigrams, it showed improved results in scoring feature set in X dataset	ET=SVM GT=LogR WT=Bayes

18.	[24]	2018	SVM, Logistic regression, LSTM+2D TF-IDF, CNN +2D TF-IDF, LSTM+ EMBEDDING, CNN + EMBEDDING	–	Micro-Area Under Curve, Macro-Area Under Curve, Accuracy	Using 2-d TF-IDF and with CNN model showed improvement	CNN+2D TF-IDF
19.	[25]	2018	CNN, LSTM	10	Area Under Curve, Receiver Operating Characteristic	Combined CNN at character level and LSTM at word level, showed improved performance over other models.	proposed hybrid classifier
20.	[26]	2017	Naïve Bayes, Decision Tree, Random Forest, NN, CBoW	10	Precision, Recall, ROC, RMSE, Cohen's Kappa ,	understanding the characteristics of bullies and aggressors on X,	RF
21.	[27]	2016	Random Forest, CNN-non-static, CNN (POS)	10	F Score	Random Forest showed better results than CNN.	Graph Based Classifier
22.	[28]	2012	Naïve Bayes, L-SVM (linear), R-SVM, (RBF), Logistic Regression, CRF	5	Accuracy, Precision, Recall, F Score	Identified key issues and framed them as Natural Language Processing task	CRF

23.	[29]	2012	Bag of Words, PLSA, Bayes method, SVM	10	Accuracy	Presented sentiment analysis technique to detect cyberbullying	LibSVM
24.	[30]	2011	Bully Tracer	–	Recall	Designed BullyTracer to detect cyberbullying instances in chat room conversation	Handmade Dictionary Based Rules
25.	[31]	2015	SVM, Logistic Regression	–	Precision, Recall, Area Under Curve, Accuracy	Increased Accuracy by 4% using the proposed model.	Logistic Regression
26.	[32]	2013	SVM	10	Precision, Recall, F score	Improved cyberbullying detection accuracy by including contextual data of user activity	SVM
27.	[33]	2012	Bag of Word, Naïve Bayes, SVM, J48, Jrip	–	Accuracy, F score,	Proposed model to detect and differentiate cyberbullying on three categories, users' intelligence, sexuality and race.	Jrip, LinearSVM
28.	[34]	2016	Logistic Regression	5	Precision, Recall, F Score, False	Showed high performance in predicting cyberbullying.	Logistic Regression

					Positive, ROC, Area Under Curve		
29.	[35]	2014	Bagging, J48, SMO, Dagging, naïve bayes, ZeroR	10	Receiver Operating Characteristic	Used social relationship as a factor to detect cyberbullying	Dagging
30.	[36]	2015	Naïve Byes, SVM, Logistic Regression		Precision, Recall, Accuracy	Showed collaborative detection outperforms the stand-alone models.	Collaborative Paradigm
31.	[37]	2013	Session-based ensemble classifier one-class classification	10	Precision, Recall, F Score	The author presented an ensemble technique to detect cyberbullying instance	Ensemble Classifier
32.	[38]	2016	Proposed Method	10	Precision, Recall	Proposed a novel model for detection of cyberbullying on internet	Proposed Method
33.	[6]	2011	Decision Tree, kNN, SVM	10	Confidence	Included personality traits related to cyberbullying and showed it affects the presence of bullying content	J48

34.	[39]	2018	Fuzzy Fingerprints, Logistic Regression, Multinomial Naïve Bayes, SVM	10	Precision, Recall, F Score	Fuzzy Fingerprint outperformed the model approach in detecting cyberbullying	Fuzzy Fingerprints
35.	[40]	2016	ZeroR, naive fusion, late fusion		Recall, Precision Accuracy, F-Score	Proposed a novel probabilistic information fusion framework	Proposed Method
36.	[41]	2016	SVM, Naïve Bayes, Logistic Regression	–	Recall, Precision, F Score, Accuracy	Author designed a system to automatically grade the instance of cyberbullying and provide appropriate response in its effect	SVM
37.	[42]	2015	SVM	–	Precision, F score, Recall,	Author showed the detection of fined grained cyberbullying categories	SVM
38.	[43]	2016	PCNN, CNN	5 and 10	Accuracy, Recall, Precision, F Score, True Positive, False Positive, True Negative, False Negative	Proposed PCNN model for cyberbullying detection, it showed better results than existing models.	PCNN

39.	[44]	2017	smSDA, mSDA, SVM, LSA, LDA, Bag of Words, sBoW,	10	Accuracy, F Score	Proposed smSDA for textual cyberbullying detection, this showed better results.	mSDA
40.	[45]	2019	Continuous BoW Semantic-enhanced BoW Model, LDA, tf-idf, SVM	5	Precision, Recall, F Score	Proposed EboW for Cyberbullying that produced improved results.	EBoW
41.	[46]	2016	SMOTE, Naïve Bayes, SVM Random Forest, and kNN	10	Precision, Recall, Area Under Curve, F Score	Random forest using SMOTE showed better results.	Random forest SMOTE only
42.	[47]	2012	SVM	–	Precision, Recall, F Score	Author used gender specific feature to improves the discrimination capacity of a classifier to detect cyberbullying.	SVM
43.	[48]	2018	SVM, Logistic Regression, CNN, CNN-LSTM,	10	Precision, Recall, F Score	Author showed that deep learning model with word embedding outperform baseline classifiers	C-LSTM
44.	[49]	2016	GHSOM, C4.5, SVM, Naïve Bayes	10	Precision, Recall, Accuracy,	Author developed a model inspired by Growing	GHSOM

					F Score, True Positive, Area Under Curve	Hierarchical SOMs and tuned it work on SM	
45.	[50]	2016	Random Forest, SMO, Mutlilayer Perceptron, J48	10	Precision, Recall, F Score	Author incorporates latent or hidden variables with supervised learning to determine potential bullying cases.	RF
46.	[51]	2020	BLSTM, GRU, LSTM, RNN	–	Recall, F Score Precision, Accuracy, AUC, True Positive, True Negative, False Positive, False Negative	Showed the empirical analysis of deep learning model for detection of cyberbullying instances	BLSTM
47.	[52]	2020	Proposed Model	–	Accuracy, Precision, Recall, F Score, True Positive, True Negative, False	The author proposed an algorithm which uses combination of textual feature to detect CB.	Propose d Method

					Positive, False Negative		
48.	[53]	2019	Raw, DeepWalk, Node2vec, GraRep, Xbully with Random Forest, LSVM, Logistic Regression and Bully, SCID	–	F Score	Author showed the proposed model Xbully out performs the state-of-the-art CB detection models	RF Xbully
49.	[54]	2020	SVM, Logistic Regression, CNN, RNN+LSTM, BiLSTM, BERT	5	F Score, p value	Proposed a novel approach to detect CB using BERT	BERT
50.	[55]	2020	ALBERT, SVM, CNN, CNN+GRU, BERT+CNN, GPT-2, n-gram,	–	Accuracy, Precision, Recall, F Score, Area Under Curve	Author proposed model showcase the high scores in multiple benchmarks and achieved an F1 score of. 95	ALBERT
51.	[56]	2018	Logistic Regression, RNN, LSTM+GLoVe+GBDT, AUTH, AUTH+LR,	10	Precision, Recall, F Score	Author incorporates community-based profiling features of X users to detect cyberbullying instances	LR + AUTH

			AUTH+HS, AUTH+WS				
52.	[57]	2017	Logistic Regression, MLP	15	AUC, Spearman		MLP+ED
53.	[58]	2020	NN+TF-IDF, NN+Word2 Vec, LSTM, LSTM + context2vec (NN), LSTM + context2vec	NA	Accuracy, Precession, Recall, F Score	Author presented model for extracting feature from images and combining it with neural network to identify new features	LSTM + context2 vec (NN combiner)
54.	[59]	2021	KNN, Naïve Bayes, SVM, ConvNet	10	Precision, Recall, Accuracy, ROC, Area Under Curve	The Presented a multi modal model for cyberbullying detection	ConvNet
55.	[60]	2021	SVM, Logistic Regression, Naïve Bayes, Random Forest, CNN, LSTM	–	Accuracy, Precision, Recall, F Score	Presented code-mixed language cyberbullying corpus against children and women	BERT+ CNN+GRU+Capsule
56.	[61]	2020	CNN, LSTM, VGG16, VGG19, Resnet50	–	Accuracy, Precision, Recall	Using VGG-16 and CNN extracted features from text and image.	VGG-16, CNN, GA
57.	[62]	2019	CNN	–	Accuracy, recall, Precision	Proposed single characterization of image and text to remove the	CNN

						need of different modules for each.	
58.	[63]	2022	SVM, DT(J45), Bi-LSTM, Logistic Regression, Multi-Layer Perceptron, Random Forest, Naïve Bayes	–	Precision, Recall, F-Score, Specificity, MSE, ROC	Used lexical meaning of text along with word order to improve accuracy.	Propose d Model
59.	[64]	2022	Elamn RNN with DEA, BiLSTM, RNN, SVM, Naïve Bayes, Random Forest	–	Accuracy, Precision, Recall, F-Score, Specificity	Used Elman RNN along with DEA (Dolphin Echolocation Algorithm) to detect Cyberbullying	E-RNN
60.	[65]	2023	SVM, TF-IDF, BOW	–	Accuracy, Precision, Recall, F1	Used SVM TF-IDF and BOW on Arabic dataset on cyberbullying.	SVM-TF-IDF
61.	[66]	2018	Random Forest Word2vec, Glove	–	Area Under Curve, Precision	Used data extracted from Reddit to train model using Word2vec skip gram model. Model showed better result than four pre trained model	Random Forest

TABLE 1.2 DETAILED SUMMARY OF THE STUDIES

3.2 RQ2: WHAT IS THE SOURCE, TYPE AND DOMAIN OF DATA THAT IS BEING UTILIZED IN CYBERBULLYING DETECTION?

The source, type and domain of data that is being used along with its language, inter-rater reliability and annotator expertise is presented in Table V. Following a review of the literature, it was discovered that the researchers used a wide range of sources to collect data, which also included publicly available information. The datasets under consideration were mostly publicly available, or were a set of arbitrary data collected in from different social media platforms such as X, YouTube, Facebook, Instagram, Wikipedia, Vine, Myspace, Reddit and so on. The presented studies made use of public datasets in addition to them. One of the data for the content analysis workshop was provided by "Fundacion Barcelona Media (FBM)", FBM made five datasets available. Among the others, authors of [14], [29] used Kongregate, Slashdot, and Myspace. Several studies were carried out using data obtained from social media sites via their APIs. The corpora retrieved covered a wide variety of cases, themes, and time durations. Detailed report on the dataset like social media from which data was collected, type of data in dataset, domain of dataset expertise of the annotator, language of dataset, etc., can be found on Table V

Social Media	No. of Studies
Kaggle	4
X	31
Myspace	8
YouTube	6
Facebook	3
FormSpring	9
Perverted Justice Website	1
Vine	3
Ask.fm	3
Slashdot	3
Reddit	1
Instagram	8

TABLE 1.3 SOURCE OF DATASET

Randomized datasets are used in a number of other studies [6]–[8], [12], [14], [16], [20]–[25], [27], [28], [30], [31], [33]–[39], [41]–[51], [53]–[59], [63], [65], [66].

Type of Media	No. of Studies
Textual	57
Visual	02
Sound	00
Picture	05
Infographic	01
Multi-Modal	05

TABLE 1.4 TYPE OF DATA IN CYBERBULLYING DETECTION

The use of random data sets on general issues from many disciplines has been proposed for research evaluations, particularly since 2011. Table III also shows the extensive list of Social Media sites that were included in the analysis of research involving the use of Cyberbullying detection. Table IV lists the different types of multimedia for
which were used to develop Cyberbullying detection models. There are no standardized datasets for detecting cyberbullying (Table III). Despite the fact that the majority of research collect data from the social networks

(e.g., YouTube, X and Facebook), the datasets are created independently by scraping the website or using a publicly available API. As a result, comparing the data is impossible. X is heavily favored social media for gathering data on cyberbullying. Formspring is a frequently used dataset that has undergone changes over the years. Formspring began with nearly 4000 samples [6], but its size has more than tripled since then [39]. Many authors only used datasets from Slashdot, Myspace, and Kongregate, which are available in various types.

Table V shows that the many of the datasets are significantly uneven and skewed, with the majority of studies using datasets with less than 25% of available samples classified as cyberbullying. This mismatch can reduce the prediction capability of a machine learning classifier. Because cyberbullying is a naturally unbalanced phenomenon in terms of its occurrence, some studies [40], [43], [46] have used artificial over sampling or under sampling methods to create a more balanced dataset, this artificial under and oversampling have shown better results in detection of cyberbullying. Non-normal distributions are likely in cyberbullying research. The

small number of cyberbullying examples on social media, in comparison to the vast majority of other types of posts, reflects this disparity in the data.

Table V also show that in most of the studies the language of dataset is in English, this represent a limitation on understanding of issue of cyberbullying on a global scale. Expanding datasets to include diverse languages is essential to create a more comprehensive, inclusive, and accurate understanding of cyberbullying behaviours worldwide, especially in multilingual and non-English-speaking regions where online harassment may differ in context and impact. In conclusion, we argue that the datasets available impede research conducted in this field, as a result, it is crucial to implement a fundamental change in how data actually represents cyberbullying so that subsequent years can be dedicated to more in-depth research.

S.no	**Social Media**	**Type of data**	**Dataset Description**	**Dataset Domain**	**Annotator expertise**	**Inter-rater Reliability**	**Language**	**Data Balance**
1.	Formspring, X, Wikipedia	Txt	12000 post form FormSpring 16000 from X 100000 from Wikipedia	Random	Three Amazon Mechanical Turk workers for FormSpring. One for X. 10 annotator via Crowd-work for Wikipedia	–	English	–

2.	YT	Txt	54,050 comments from 3825 users	Random	2 Graduate Student	93 %	English	–
3.	Okey (CCSoftOkey Player Abuse (COPA) Database)	Txt	6 months record for player interaction of over 80,000 Okey games	Player information, chat log and complaints	Player Reported	–	–	–
4.	Perverted-justice website	Txt	Random	Romantic movies, chats	–	–	–	–
5.	Formspring.me and Myspace	Txt	Random	–	–	–	–	–
6.	Formspring.me	Txt	13.652 post from formspring.me	Random	3 Amazon's Mechanical Turk	–	English	–
7.	Perverted-Justice (PJ)	Txt	–	Sexual Conversation	Manually labelled by Human Expert	–	–	–
8.	Myspace, Kongregate,	Txt	Fundacion Barcelona Media	Random	–	–	–	–

	and Slashdot.							
9.	Vine	Vis	652,000 media from Vine	Well Known Celebrities	5 CrowdFlower	76.6% and 79.49 %	English	–
10.	X	Txt	Tweets via customized crawler written in Python.	Random	–	–	English	–
11.	Vine	Vis	59,560 users and 652000 media sessions from vine.	User profiles	5 CrowdFlower	76.6 and 79.49 %	English	–
12.	X	Txt	650000 tweets 1 M baseline tweets	Random and hate related	5 CrowdFlower	0.2189	English	0.1115
13.	X	Txt	18,504 tweets	School students and staff members	–	–	–	–
14.	Ask.fm	Txt	Collected user profile information from Ask.fm based key	Random	12 graduate students in	0.8125	English	–

			term which relate to cyberbullying		Computer Science			
15.	Ask.fm, Insta, and X	Txt	296,308 X tweets, 2,863,801 Ask.fm Q/A, 9,828,760 messages from Insta.	Random	Amazon Mechanical Turk	–	English	–
16.	X	Txt	91,431 tweets from X	Random	Manually labelled	–	Arabic	0.062
17.	X, Wikipedia Talk pages	Txt	15,979 tweets, 11,304 Wikipedia Talk pages with annotated comments	Random	Crowd Source	–	English, German	ET=.32 GT=.22 WT-=.22
18.	X	Txt	Random aggressive comments from X.	Random	–	–	English	–
19.	Kaggle	Txt	8815 comments from Kaggle.	Random	–	–	English	0.32
20.	X	Txt	650000 tweets	Random and hate related	Crowd Source	0.54	English	0.08

			1 M baseline tweets					
21.	4chan.org, 2ch.hk	Txt	1000 English messages	Random	User reported confirmed by moderators	–	English	0.28
22.	X	Txt	990 tweets from X	Random	NA	–	English	0.39
23.	Kongregate, Slashdot and Myspace	Txt	Data from the workshop on Content Analysis	Games	NA	–	English	–
24.	MySpace1	Txt	unspecified	Random	3 Undergrad. Research assistants	–	English	–
25.	Kaggle	Txt	2647 comments	Random	–	–	English	0.272
26.	YT	Txt	4626 comments from 3858 users	Movies	manually labelled	–	English	0.097
27.	YT	Txt	1500 YT comments	Random	Three annotators of whom one was an educator	0.4	English	–

28.	Instagram	Images	41k Insta user ids	personalities, like actors, actress, singers, dancer, etc	five CrowdFlower	–	English	0.29
29.	X	Txt	CWA 2.0 dataset which contained 900,000 posts of 27135 users and selected 4865 messages of 2150 pair of users	Random	3 students	0.93	English	0.019
30.	X	Txt	Public available X dataset	Random	Manually labelled	–	English	0.152
31.	MySpace2, X, ,Slashdot (CAW 2.0) and Kongregate	Txt	Author Used dataset provided by Fundacion Barcelona Media. Data was collected from X, Myspace, Kongregate, and Slashdot	Random	–	–	English	–

32.	School board Bulletins (BBS)	Txt	The author collected 2222 entries from informal school bulletin boards (BBS).	Random	Unknown experts (Internet Patrol)	–	Japanese	0.128
33.	Formspring.me	Txt	18,554 users information from Formspring.me., 2696 posts for training, 1219 for testing.	Random	3 Amazon Mechanical Turk	–	English	0.142
34.	Formspring4	Txt	13160 labelled texts from Fromspring.me during summer 2010	Random	3 Amazon Mechanical Turk	–	English	0.194
35.	X CAW 2.0	Txt	4865 tweets from X	Random	three labellers	–	English	0.186
36.	Train (Formspring and Myspace); Test (X)	Txt	3279 text data from X, myspace and FormSpring	random	3 Undergrad. Research assistants	–	English	0.12

37.	AskFM	Txt	85485 post form askFM	random	2 annotator	0.69	Dutch	0.067
38.	X and Formsprig.me	Txt	1313 post from X and 13,000 from Formspring.me	Random	3 Amazon Mechanical Turk	–	English	0.0698
39.	X and Myspace	Txt	7321 tweets from X 1539 from Myspace.	Random	Manually labelled	–	English	0.28
40.	X	Txt	1762 tweets on 6 Aug 2011.	Random	Manually labelled	–	English	0.39
41.	X	Txt	2.5 million tweets geo-tagged between January 2015 and February 2015.	Random	three experts	–	English	0.06
42.	Myspace	Txt	381,000 posts in about 16,000 threads Selected about 2200 post	Random	manually labelled by three students	–	English	–
43.	Formspring4	Txt	Author collected 13160 labelled texts from	Random	3 Amazon Mechanical Turk	–	English	0.194

			the site in summer 2010					
44.	Formspring.me, YT, and X	Txt	Author collected post from FormSpring, X and YT	Random	3 Amazon Mechanical Turk	–	English	0.06 FS 0.14 YT
45.	X	Txt	2500 public Tweets using the TAGS archiving tool and selected 1689 instances	Random	2 Annotator	0.833.	English	0.22
46.	Kaggle	Txt	Dataset for Kaggle repository	Random	NA	–	English	–
47.	FB	Txt	202 Random entries from FB	Melania Trump's FB post comments	Manually annotated	0.76073.	English	0.7
48.	Instagram and Vine	Txt, Picture	Author collected 3188 session from Insta and vine	Random	–	–	English	0.31
49.	Formspring, X, and Wikipedia	Txt	Author collected data. From X, FormSpring	Random	–	–	English	T=0.633. W=.443.

			and Wikipedia					F=.368
50.	X	Txt	Author collected data from X	Random	Crowdsourcing	–	English	0.06
51.	X	Txt	16907 Tweets using Tweepy	Random	Manually Annotated	0.84	English	0.317
52.	Wikipedia	Txt	63M comments from discussion from Wikipedia and 115737 annotated comments	Random	CrowdFlower	0.45	English	0.117
53.	Insta, X	Txt Picture	Collected post from Insta and X	Random	–	–	English	I=.3 T=NA
54.	YT, Insta and X	Txt Picture Infographic	10000 comments from YT Insta X	Random	–	–	NA	–
55.	X	Txt	5062 Tweets	Women and children	Two human annotators having linguistic background	0.85	Hinglish	–

56.	X, Instagram, FB,	Txt, Picture	2100 post from FB, X and Insta	Keywords based on public shamming	Two Experts	0.86	English	0.7
57.	X, Instagram, FB,	Txt, Picture	2100 post from FB, X and Insta.	Keywords based on public shamming	Two Experts	0.86	English	0.7
58.	Instagran	Txt	10421 comments from 2218 session	Random	–	–	English	0.21
59.	X	Txt	10000 comments from X, over sample to 13016 comments.	32 cyberbullying keywords	Three Experts	0.91	English	1
60.	X, YT	Txt	30000 comments from YT and X	Random	–	–	Arabic	–
61.	Reddit, Kaggle	Txt	245,292,434 reddit comment used for word2vec, 6,594 comments from Kaggle dataset	Random	–	_	English	0.75

TABLE 1.5 DETAILED SUMMARY OF DATASET USED IN STUDIES

3.3 RQ3: WHAT GUIDELINES FOR FUTURE CYBERBULLYING DETECTION RESEARCH MAY BE DERIVED FROM THE FINDINGS OF THE REVIEW?

- **New models and potential for Improvement:** Ability to detect and forecast cyberbullying behaviour is a difficult task, with anonymised cyberbullying adding to the difficulty. Despite the fact that researchers are eager to implement diffrent approaches, only very few approaches have been investigated. Neural network, deep learning, evolutionary computing, ensemble methods, and hybrids model such as, fuzzy based models have garnered a lot of attention in Cyberbullying detection. More research is needed to improve Cyberbullying detection using new pre-trained models such as BERT, ALBERT, and others; very few studies have used them, and their full potential has yet to be explored.
- Use of multi-modal data: Majority studies focus on single modality of data i.e. textual, visual, audio or image but very few studies have used multi modal data for their studies, cyberbullying doesn't always incorporate single mode the perpetrator uses various means to harass and oppress its victims, it often seen the perpetrator uses various form of media to target their victims. The majority of data shared on social networks are multimedia data (e.g., photographs with captions, live video streaming, and audio files) .These data are absent due to the scarcity of datasets, including image and video data. Using multi-modal data can provide the enhanced understanding of the context in which the bullying occurs.
- **Improving data collection process:** The process of data collection is often times done based on keywords or user profile. Using keywords to track post have some keywords often introduce bias, which can hamper the performance of the cyberbullying model and provide with biased outcome. The best way to overcome this issue is use wide range of post and domain from which the data is collected.

  The absence of information regarding the construction of datasets is major issue. The majority of research does not include criteria for labelling the data samples. Similarly, inter-rater dependability and the competence of annotators are sometimes overlooked. This issue reflects the absence of a uniform definition of cyberbullying and shared, accessible data annotation techniques, as previously noted. As a result, the variability of cyberbullying definitions and the criteria used to classify text material as cyberbullying greatly raise the danger of interrater subjectivity biases.

Class imbalance is another issue that can hamper the quality of data and affect the performance of the model. Most of the datasets for cyberbullying are naturally imbalanced. In the real world, the cases of Cyberbullying are significantly low compared to non-cyberbullying posts; this causes a class imbalance problem in the dataset, it can also be seen in Table I most of the datasets are imbalanced. To overcome this issue, there are many techniques that can be used, like class weighting, resampling, and data augmentation (SMOTE).

- **Using appropriate Key Performance Indicator:** Precision, recall and F-Score are most commonly used key performance indicators as seen in Table II. Selection of which key performance indicator to use is important. The selection depends on nature of dataset being used; selection inappropriate performance indicator may result in better performance according to selected performance indicator. This might mislead the researcher to find the results improved, but in reality, this result is misleading and doesn't reflect the reported better performance. For example, a dataset containing severe class imbalance is used to detect the cyberbullying. The model being imbalance tends to identify instance as majority class, this leads to high accuracy, since most instance of data is non-cyberbullying. So, if we use accuracy as performance measure, then accuracy for this will be high. Even though the model is successful in detecting non cyberbullying instances but it weak at identifying the minority class i.e., cyberbullying instances, which is the major task of the model. Hence, using accuracy alone in this imbalanced scenario provides a skewed perspective, as the model's performance is heavily biased towards the majority class.

Special thought must be given when selection the key performance indicators so as to ensure that performance reported reflect the true capabilities of the model.

- **Including Personal, Psychological and Social-Cultural features:** An individual's conduct and sentiments regarding cyberbullying are substantially influenced by personal characteristics. These variables are unique to each individual and often result in distinct beliefs and behaviors related to cyberbullying. Many researchers have studied the association between gender, race, age, and cyberbullying. Gender, age, and race are significant factors that influence the commission and experience of cyberbullying, resulting in detrimental outcomes for both the perpetrator and victim. There are many psychological features that either precede or follow cyberbullying activity, or both. These characteristics include traits of personality, tension, anxiety, depression, emotional maturity, vengeance, solitude, irritation, self-esteem, aggression, empathy, antisocial conduct, insecurity, internalizing behaviour and

jealousy. These psychological variables are crucial for explaining the phenomena of cyberbullying. Understanding these psychological aspects can contribute to a thorough knowledge of the complex nature of cyberbullying behaviour.

4. CONCLUSION

Since the rise in number of cyberbullying cases there is high need for a computational model to identify cyberbullying. It necessitates the search for models that integrate soft computing approaches' intellect, self-tuning, and comprehension with fields such as natural language processing, cognitive and behavioral science, and artificial intelligence. Potential victims of cyberbullying can be saved by early detection of cyberbullying with the use of machine learning and deep learning models Moreover, it can be supplemented by warning interfaces that notifies the bully of their action or delay their action this encourage bully self-reflect on their action. It is also important that these models are able to distinguish cyberbullying form non cyberbullying instances which may appear similar, such as use of profanity among peer in humorous way.

One of the risks of inaccurate cyberbullying detection is that it can reduce the responsiveness of these solutions, causing users to disregard them. Another concern associated with erroneously recognizing cyberbullying is the development of poor digital solutions that are incapable of preventing and intervening in genuine cyberbullying incidents, thereby failing to safeguard users from potentially dangerous scenario. To overcome these issues, the standard of future classifiers for the identification of cyberbullying must be enhanced. Also focus should be made on detecting Cyberbullying in multilingual and multimodal scenarios so as to reach wider audience across the globe.

5. REFERENCES

[1] C. Langos, “Cyberbullying: The Challenge to Define,” *Cyberpsychology, Behavior, and Social Networking*, vol. 15, no. 6, pp. 285–289, Jun. 2012, doi: 10.1089/cyber.2011.0588.

[2] “Facebook Quaterly Earning.” https://s21.q4cdn.com/399680738/files/doc_financials/2021/q3/FB-Earnings-Presentation-Q3-2021.pdf

[3] S. Hinduja and J. W. Patchin, “Cyberbullying: An Exploratory Analysis of Factors Related to Offending and Victimization,” *Deviant Behavior*, vol. 29, no. 2, pp. 129–156, Jan. 2008, doi: 10.1080/01639620701457816.

[4] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan, “An Introduction to MCMC for Machine Learning,” *Machine Learning*, vol. 50, no. 1, pp. 5–43, Jan. 2003, doi: 10.1023/A:1020281327116.

[5] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, “Detection of Harassment on Web 2.0,” in *Content Analysis in the WEB 2.0 (CAW2.0) Workshop at WWW2009*, Madrid, Spain, Apr. 2009.

[6] K. Reynolds, A. Kontostathis, and L. Edwards, “Using Machine Learning to Detect Cyberbullying,” in *2011 10th International Conference on Machine Learning and Applications and Workshops*, Dec. 2011, pp. 241–244. doi: 10.1109/ICMLA.2011.152.

[7] S. Agrawal and A. Awekar, “Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms,” in *Advances in Information Retrieval*, G. Pasi, B. Piwowarski, L. Azzopardi, and A. Hanbury, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 141–153. doi: 10.1007/978-3-319-76941-7_11.

[8] M. Dadvar, D. Trieschnigg, and F. de Jong, “Experts and Machines against Bullies: A Hybrid Approach to Detect Cyberbullies,” in *Advances in Artificial Intelligence*, M. Sokolova and P. van Beek, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014, pp. 275–281. doi: 10.1007/978-3-319-06483-3_25.

[9] K. Balci and A. A. Salah, “Automatic analysis and identification of verbal aggression and abusive behaviors for online social games,” *Computers in Human Behavior*, vol. 53, pp. 517–526, Dec. 2015, doi: 10.1016/j.chb.2014.10.025.

[10] D. Michalopoulos, I. Mavridis, and M. Jankovic, “GARS: Real-time system for identification, assessment and control of cyber grooming attacks,” *Computers & Security*, vol. 42, pp. 177–190, May 2014, doi: 10.1016/j.cose.2013.12.004.

[11] B. S. Nandhini and J. I. Sheeba, “Online Social Network Bullying Detection Using Intelligence Techniques,” *Procedia Computer Science*, vol. 45, pp. 485–492, Jan. 2015, doi: 10.1016/j.procs.2015.03.085.

[12] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards, “Detecting cyberbullying: query terms and techniques,” in *Proceedings of the 5th Annual ACM Web Science Conference*, in WebSci ’13. New York, NY, USA: Association for Computing Machinery, May 2013, pp. 195–204. doi: 10.1145/2464464.2464499.

[13] N. Potha, M. Maragoudakis, and D. Lyras, “A biology-inspired, data mining framework for extracting patterns in sexual cyberbullying data,” *Knowledge-*

Based Systems, vol. 96, pp. 134–155, Mar. 2016, doi: 10.1016/j.knosys.2015.12.021.

[14] V. Nahar, S. Al-Maskari, X. Li, and C. Pang, "Semi-supervised Learning for Cyberbullying Detection in Social Networks," in *Databases Theory and Applications*, H. Wang and M. A. Sharaf, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014, pp. 160–171. doi: 10.1007/978-3-319-08608-8_14.

[15] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, and S. A. Mattson, "Careful what you share in six seconds: Detecting cyberbullying instances in Vine," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, in ASONAM '15. New York, NY, USA: Association for Computing Machinery, Aug. 2015, pp. 617–622. doi: 10.1145/2808797.2809381.

[16] G. Sarna and M. P. S. Bhatia, "Content based approach to find the credibility of user in social networks: an application of cyberbullying," *Int. J. Mach. Learn. & Cyber.*, vol. 8, no. 2, pp. 677–689, Apr. 2017, doi: 10.1007/s13042-015-0463-1.

[17] R. I. Rafiq, H. Hosseinmardi, S. A. Mattson, R. Han, Q. Lv, and S. Mishra, "Analysis and detection of labeled cyberbullying instances in Vine, a video-based social network," *Soc. Netw. Anal. Min.*, vol. 6, no. 1, p. 88, Sep. 2016, doi: 10.1007/s13278-016-0398-x.

[18] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Detecting Aggressors and Bullies on Twitter," in *Proceedings of the 26th International Conference on World Wide Web Companion*, in WWW '17 Companion. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, Apr. 2017, pp. 767–768. doi: 10.1145/3041021.3054211.

[19] C. R. Sedano, E. L. Ursini, and P. S. Martins, "A Bullying-Severity Identifier Framework Based on Machine Learning and Fuzzy Logic," in *Artificial Intelligence and Soft Computing*, L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017, pp. 315–324. doi: 10.1007/978-3-319-59063-9_28.

[20] Z. Ashktorab, E. Haber, J. Golbeck, and J. Vitak, "Beyond Cyberbullying: Self-Disclosure, Harm and Social Support on ASKfm," in *Proceedings of the 2017 ACM on Web Science Conference*, in WebSci '17. New York, NY, USA: Association for Computing Machinery, Jun. 2017, pp. 3–12. doi: 10.1145/3091478.3091499.

[21] E. Raisi and B. Huang, "Cyberbullying Detection with Weakly Supervised Machine Learning," in *2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Aug. 2017, pp. 409–416.

[22] B. Haidar, M. Chamoun, and A. Serhrouchni, "Multilingual cyberbullying detection system: Detecting cyberbullying in Arabic content," in *2017 1st Cyber Security in Networking Conference (CSNet)*, Oct. 2017, pp. 1–8. doi: 10.1109/CSNET.2017.8242005.

[23] P. Bourgonje, J. Moreno-Schneider, A. Srivastava, and G. Rehm, "Automatic Classification of Abusive Language and Personal Attacks in Various Forms of Online Communication," in *Language Technologies for the Challenges of the Digital Age*, G. Rehm and T. Declerck, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 180–191. doi: 10.1007/978-3-319-73706-5_15.

[24] J. Chen, S. Yan, and K.-C. Wong, "Verbal aggression detection on Twitter comments: convolutional neural network for short-text sentiment analysis," *Neural Comput & Applic*, vol. 32, no. 15, pp. 10809–10818, Aug. 2020, doi: 10.1007/s00521-018-3442-0.

[25] "A Hybrid Deep Learning System of CNN and LRCN to Detect Cyberbullying from SNS Comments | SpringerLink." https://link.springer.com/chapter/10.1007/978-3-319-92639-1_47

[26] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean Birds: Detecting Aggression and Bullying on Twitter." arXiv, May 12, 2017. doi: 10.48550/arXiv.1702.06877.

[27] D. Gordeev, "Detecting State of Aggression in Sentences Using CNN," in *Speech and Computer*, A. Ronzhin, R. Potapova, and G. Németh, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 240–245. doi: 10.1007/978-3-319-43958-7_28.

[28] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, in NAACL HLT '12. USA: Association for Computational Linguistics, Jun. 2012, pp. 656–666.

[29] V. Nahar, S. Unankard, X. Li, and C. Pang, "Sentiment Analysis for Effective Detection of Cyber Bullying," in *Web Technologies and Applications*, Q. Z. Sheng, G. Wang, C. S. Jensen, and G. Xu, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2012, pp. 767–774. doi: 10.1007/978-3-642-29253-8_75.

[30] J. Bayzick, A. Kontostathis, and L. Edwards, “Detecting the Presence of Cyberbullying Using Computer Software,” 2011.

[31] V. S. Chavan and S. 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)*, Aug. 2015, pp. 2354–2358. doi: 10.1109/ICACCI.2015.7275970.

[32] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, “Improving Cyberbullying Detection with User Context,” in *Advances in Information Retrieval*, P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, and E. Yilmaz, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2013, pp. 693–696. doi: 10.1007/978-3-642-36973-5_62.

[33] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, “Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying,” *ACM Transactions on Interactive Intelligent Systems*, vol. 2, Sep. 2012, doi: 10.1145/2362394.2362400.

[34] H. Hosseinmardi, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, “Prediction of cyberbullying incidents in a media-based social network,” in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Aug. 2016, pp. 186–192. doi: 10.1109/ASONAM.2016.7752233.

[35] Q. Huang, V. K. Singh, and P. K. Atrey, “Cyber Bullying Detection Using Social and Textual Analysis,” in *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*, in SAM ’14. New York, NY, USA: Association for Computing Machinery, Nov. 2014, pp. 3–6. doi: 10.1145/2661126.2661133.

[36] A. Mangaonkar, A. Hayrapetian, and R. Raje, “Collaborative detection of cyberbullying behavior in Twitter data,” in *2015 IEEE International Conference on Electro/Information Technology (EIT)*, May 2015, pp. 611–616. doi: 10.1109/EIT.2015.7293405.

[37] V. Nahar, X. Li, C. Pang, and Y. Zhang, “Cyberbullying Detection based on text-stream classification,” in *Australasian Data Mining Conference*, 2013.

[38] M. Ptaszynski *et al.*, “Sustainable cyberbullying detection with category-maximized relevance of harmful phrases and double-filtered automatic optimization,” *International Journal of Child-Computer Interaction*, vol. 8, pp. 15–30, May 2016, doi: 10.1016/j.ijcci.2016.07.002.

[39] H. Rosa, J. P. Carvalho, P. Calado, B. Martins, R. Ribeiro, and L. Coheur, “Using Fuzzy Fingerprints for Cyberbullying Detection in Social Networks,”

in *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Jul. 2018, pp. 1–7. doi: 10.1109/FUZZ-IEEE.2018.8491557.

[40] V. K. Singh, Q. Huang, and P. K. Atrey, "Cyberbullying detection using probabilistic socio-textual information fusion," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Aug. 2016, pp. 884–887. doi: 10.1109/ASONAM.2016.7752342.

[41] R. Sugandhi, A. Pande, A. Agrawal, and H. Bhagat, "Automatic Monitoring and Prevention of Cyberbullying," *International Journal of Computer Applications*, vol. 144, no. 8, pp. 17–19, Jun. 2016.

[42] C. V. Hee *et al.*, "Detection and Fine-Grained Classification of Cyberbullying Events," in *Recent Advances in Natural Language Processing*, 2015.

[43] X. Zhang *et al.*, "Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network," in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2016, pp. 740–745. doi: 10.1109/ICMLA.2016.0132.

[44] R. Zhao and K. Mao, "Cyberbullying Detection Based on Semantic-Enhanced Marginalized Denoising Auto-Encoder," *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 328–339, Sep. 2017, doi: 10.1109/TAFFC.2016.2531682.

[45] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," in *Proceedings of the 17th International Conference on Distributed Computing and Networking*, in ICDCN '16. New York, NY, USA: Association for Computing Machinery, Jan. 2016, pp. 1–6. doi: 10.1145/2833312.2849567.

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

[47] M. Dadvar, F. de Jong, R. Ordelman, and D. Trieschnigg, "Improved Cyberbullying Detection Using Gender Information," presented at the Cognitive Processing - COGN PROCESS, Jan. 2012.

[48] H. Rosa, D. Matos, R. Ribeiro, L. Coheur, and J. P. Carvalho, "A 'Deeper' Look at Detecting Cyberbullying in Social Networks," in *2018 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2018, pp. 1–8. doi: 10.1109/IJCNN.2018.8489211.

[49] M. Di Capua, E. Di Nardo, and A. Petrosino, "Unsupervised cyber bullying detection in social networks," in *2016 23rd International Conference on*

Pattern Recognition (ICPR), Dec. 2016, pp. 432–437. doi: 10.1109/ICPR.2016.7899672.

[50] P. Nand, R. Perera, and A. Kasture, "'How Bullying is this Message?': A Psychometric Thermometer for Bullying," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 695–706. [Online]. Available: https://aclanthology.org/C16-1067

[51] C. Iwendi, G. Srivastava, S. Khan, and P. K. R. Maddikunta, "Cyberbullying detection solutions based on deep learning architectures," *Multimedia Systems*, vol. 29, no. 3, pp. 1839–1852, Jun. 2023, doi: 10.1007/s00530-020-00701-5.

[52] M. Fortunatus, P. Anthony, and S. Charters, "Combining textual features to detect cyberbullying in social media posts," *Procedia Computer Science*, vol. 176, pp. 612–621, Jan. 2020, doi: 10.1016/j.procs.2020.08.063.

[53] L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu, "XBully: Cyberbullying Detection within a Multi-Modal Context," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, in WSDM '19. New York, NY, USA: Association for Computing Machinery, Jan. 2019, pp. 339–347. doi: 10.1145/3289600.3291037.

[54] S. Paul and S. Saha, "CyberBERT: BERT for cyberbullying identification," *Multimedia Systems*, vol. 28, no. 6, pp. 1897–1904, Dec. 2022, doi: 10.1007/s00530-020-00710-4.

[55] J. K. Tripathy, S. S. Chakkaravarthy, S. C. Satapathy, M. Sahoo, and V. Vaidehi, "ALBERT-based fine-tuning model for cyberbullying analysis," *Multimedia Systems*, vol. 28, no. 6, pp. 1941–1949, Dec. 2022, doi: 10.1007/s00530-020-00690-5.

[56] P. Mishra, M. Del Tredici, H. Yannakoudakis, and E. Shutova, "Author Profiling for Abuse Detection," in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1088–1098. [Online]. Available: https://aclanthology.org/C18-1093

[57] E. Wulczyn, N. Thain, and L. Dixon, "Ex Machina: Personal Attacks Seen at Scale." arXiv, Feb. 25, 2017. doi: 10.48550/arXiv.1610.08914.

[58] N. Rezvani, A. Beheshti, and A. Tabebordbar, "Linking textual and contextual features for intelligent cyberbullying detection in social media," in *18th International Conference on Advances in Mobile Computing and Multimedia, MoMM2020 - Proceedings*, P. D. Haghighi, I. L. Salvadori, M. Steinbauer, I. Khalil, and G. Kotsis, Eds., in ACM International Conference Proceeding

Series. Association for Computing Machinery, Inc, 2020, pp. 3–10. doi: 10.1145/3428690.3429171.

[59] A. Kumar and N. Sachdeva, "Multimodal cyberbullying detection using capsule network with dynamic routing and deep convolutional neural network," *Multimedia Systems*, vol. 28, no. 6, pp. 2043–2052, Dec. 2022, doi: 10.1007/s00530-020-00747-5.

[60] K. Maity and S. Saha, "BERT-Capsule Model for Cyberbullying Detection in Code-Mixed Indian Languages," in *Natural Language Processing and Information Systems*, E. Métais, F. Meziane, H. Horacek, and E. Kapetanios, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 147–155. doi: 10.1007/978-3-030-80599-9_13.

[61] K. Kumari and J. P. Singh, "Identification of cyberbullying on multi-modal social media posts using genetic algorithm," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 2, p. e3907, 2021, doi: 10.1002/ett.3907.

[62] K. Kumari, J. P. Singh, Y. K. Dwivedi, and N. P. Rana, "Towards Cyberbullying-free social media in smart cities: a unified multi-modal approach," *Soft Comput*, vol. 24, no. 15, pp. 11059–11070, Aug. 2020, doi: 10.1007/s00500-019-04550-x.

[63] S. Abarna, J. I. Sheeba, S. Jayasrilakshmi, and S. P. Devaneyan, "Identification of cyber harassment and intention of target users on social media platforms," *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105283, Oct. 2022, doi: 10.1016/j.engappai.2022.105283.

[64] B. A. H. Murshed, J. Abawajy, S. Mallappa, M. A. N. Saif, and H. D. E. Al-Ariki, "DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform," *IEEE Access*, vol. 10, pp. 25857–25871, 2022, doi: 10.1109/ACCESS.2022.3153675.

[65] A. M. Alduailaj and A. Belghith, "Detecting Arabic Cyberbullying Tweets Using Machine Learning," *Machine Learning and Knowledge Extraction*, vol. 5, no. 1, pp. 29–42, Jan. 2023, doi: 10.3390/make5010003.

[66] T. Bin Abdur Rakib and L.-K. Soon, "Using the Reddit Corpus for Cyberbully Detection," in *Intelligent Information and Database Systems*, N. T. Nguyen, D. H. Hoang, T.-P. Hong, H. Pham, and B. Trawiński, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 180–189. doi: 10.1007/978-3-319-75417-8_17.