Al-Nahrain Journal of Science
ANJS, Vol.27 (1), March, 2024, pp. 121-135

A Review of the Past Decade's Research Outcomes on Classifying Cyberbullying

Zainab Haider Ameen*, Farah Saad Al-Mukhtar, Ehsan Qahtan Ahmed, Ban N. Dhannoon

Department of Computer Science, College of Science, Al-Nahrain University, Jadriya, Baghdad, Iraq

Article's Information

Received: 07.08.2023
Accepted: 31.10.2023
Published: 15.03.2024

Keywords:
Cyberbullying
Deep learning
Machine learning
Social media
Word Embedding

Abstract

Cyberbullying is a new form of online violence that has emerged as a result of the social media industry's explosive expansion since it allows for indirect communication. Despite the usefulness of digital media, it has been used by weak people to threaten and bully others online. In the last ten years, research has shown that children and teenagers are increasingly experiencing cyberbullying as a concern. This paper examines the research conducted during the previous ten years, categorizes it, and presents statistical analyses of the data collected during that time. A table is used to present various data, including the dataset that was used, its size (number of samples, posts, or messages), the methods that were employed, and the metrics that were gathered from the examined research that was taken from the publications that were looked into. This survey will be helpful to everyone who wants to advance their understanding of how machine learning may be used to identify cyberbullying, and it may help create a social media environment that is safe and relatively healthy by automatically identifying bullying communications.

http://doi.org/10.22401/ANJS.27.1.11

*Corresponding author email address: zainab.h.ameen@nahrainuniv.edu.iq

1.Introduction

Since technology has advanced, particularly in the areas of the Internet and social media platform applications, a person may know hundreds of people in different parts of the world and establish relationships with them. However, despite these benefits, there are negative aspects, including the phenomenon of cyberbullying. Cyberbullying can be defined as repeated bullying behavior over digital devices through sending, posting, or sharing harmful, negative content for those who cannot easily defend him/herself. Cyberbullying can happen through many platforms like messaging, gaming, computers, and mobile phones. Almost fifty percent of American youth are suffering from cyberbullying [1]. In the Arab world, especially in the preparatory stage, 20.9% of adolescents reported being bullied in the United Arab Emirates, while the percentage reached 31.9%, 39.1%, 44.2%, and 33.6% in Morocco, Lebanon, Oman, and Jordan, respectively[2]. These rates indicate that cyberbullying is on the rise in the Arab world. A report on cyberbullying shows that 60% of young people in Gulf countries admitted openly to cyberbullying among their peers [3]. A study by showed that only a quarter of hackers bully their victims offline [4]. This indicates that the Internet encouraged most fraudsters to bully others. In turn, they did not think of bullying head-on. Concerning cyberbullying, several national and international initiatives over the past few years to increase children's online safety like Belgian governmental initiatives, KiVa (http://www.kivaprogram.net), the 'Non-au harcèlement' campaign in France, a Finnish cyberbullying prevention program have been launched and helplines (e.g., \clicksafe.be, veiligonline.be, mediawijs.be) that provide information about online safety. Several national and international programs have been created in recent years to combat cyberbullying and improve children's online safety. The French "Non-au arcèlement" campaign, the Belgian government, the Finnish cyberbullying prevention program

KiVa (http://www.kivaprogram.net), and helplines offering online safety information include clickafe.be, veiligonline.be, and mediawijs.be. Unfortunately, despite these efforts, there is still a lot of offensive and harmful stuff online. Despite all of these efforts, a lot of undesirable and harmful content still exists online. Between 20% and 40% of teenagers encounter cyberbullying, according to researchers who looked at a range of quantitative studies on the topic [2,5]. Focusing on 12 to 17-year-olds in the US, it was found that 72% of them had experienced cyberbullying at least once in the year before the survey. 9 to 26-year-olds in Australia, the UK, the US, and Canada were polled by the authors of [6], who discovered that 29% of respondents had encountered online harassment. In a different survey, 11% of 2,000 Flemish high school students (aged 12 to 18) reported having been the victim of cyberbullying at least once in the six months before the study [7]. Finally, a thorough 2014 EU Kids Online study found that 20% of children aged 11 to 16 had access to hateful content online [8]. Additionally, young adults were 12% more likely to experience cyberbullying than in 2010, which shows that cyberbullying is a growing problem.

2. Literatures Review

An overview of previously published work on cyberbullying will be provided to place the work presented in our study within the relevant body of literature and ensure that useful context is provided. Emphasis is placed on research work published in the Direct Science Database. The organization of this section is based on the years of publication, with less recent works presented first.

Reynolds et al. 2011 employed machine learning techniques to identify cyberbullying [1]. The investigation uncovered victimization and bullying language trend. Their data was collected from a high percentage of bullying content obtained using the curated question-and-answer Formpring.me website. The data was categorized using Amazon's Turk Mechanical web service. Quadruple learning techniques (C4.5 decision tree, JRIP-based neighbor, k-nearest, and support vector machine-based function based on SMO (Sequential Minimal Optimization) to predict cyberbullying cases accurately. Vinita et al. 2012 investigated an effective technique for detecting and classifying the most influential people (predators and victims) [2]. They start with a dataset containing data collected from the Kongregate, Slashdot, and MySpace websites. The proposed discovery graph model results indicate that their method is very accurate. They introduced an effective sentiment analysis technique using PLSA (Latent Probabilistic Semantic Analysis) for feature selection to detect cyberbullying messages and used the HITS (Hyperlink Invoked Subject Search) algorithm to calculate scores and rank the most influential people. [3], investigate the SVM (support vector machine) algorithm for Sentiment Analysis on Bullying. The paper contains four steps: (1) collecting raw words, (2) collecting online documents, (3) generating feature extracts, and (4) building a text taxonomy. For eight months, the SVM applied 3,001,427 traces of bullying to Twitter posts about bullying and made several interesting discoveries. They do a standard 7-class SVM training on the Wikipedia group. This study encourages the community to devote more effort to analyzing sentiment about bullying, to reliably identify at-risk individuals. Dadvar et al. in 2014 investigated cyberbullying by identifying harassing comments instead of stopping the events by identifying the bullies [4]. On YouTube, the automatic identification of bullies was investigated. The automatic detection systems (an expert system, supervised machine learning models, and a hybrid type combining the two) were given a score reflecting the number of "bullies" in order to compare them. They demonstrate that the expert system outperforms the machine learning models, including two hybrid models that integrate the two classifiers (decision trees, and SVM with a linear kernel).

Nalini K. and Sheela L. J. Proposed in 2015 a practical approach for detecting cyberbullying messages from Twitter through a feature selection weighting scheme [5]. They employ a weighted TF-IDF (Term Frequency-Inverse Document Frequency) function, in which bullying-like characteristics are scaled by a factor of 2 to improve the classification performance. When compared to other approaches, weighted TF-IDF produced superior overall results. This summarizes their strategy to employ more deductive language in the detrimental posts. The classifier's performance was assessed using the Accuracy (Acc), Recall, and F-1 scaling based on the higher-ranked features produced by the LDA (Linear Discriminant Analysis) method against the truth set, as tested on data sets. The SVM was applied to a two-class classification problem using a linear kernel. The presence of cyberbullying words is acknowledged by Nandhini et al. and Capua et al. respectively (see reference [6,7]), where a detected approach was identified. They also classified cyberbullying acts in a social network as terrorism, harassment, and racism. The fuzzy rule set was used to gather pertinent data for classification from the input, and then the genetic algorithm was used to optimize the parameters and produce a precise output. The output from the learning unit is then used to categorize online bullying behaviors using the Naive approach. The accuracy of the proposed methodology was 87% for the MySpace dataset and 86% for Formspring.me dataset. Capua et al. automatically detected bully traces across a social network, they proposed techniques derived from Natural Language Processing (NLP) and machine learning [7]. They apply an unsupervised Growing Hierarchical Self-Organizing Map (GHSOM), Naive Bayes (NB), and SVM for classification purposes on three different datasets (Formspring.me, Youtube, and Twitter). Results indicate that their model achieves reasonable performance using GHSOM and could be usefully applied to build concrete monitoring applications to mitigate the heavy social problem of cyberbullying. Zhong et al. created early warning systems for the anticipation of posted photographs that are prone to attacks. They also examined the identification of cyberbullying in photo-sharing networks. They investigate using posted images and captions to detect bullying in response to shared content. In addition to the standard image and text features, they take advantage of several specific features from image annotations and a pre-trained convolutional neural network on image pixels. These features helped spot cyberbullying in posted comments.

Haidar et al. in 2017 developed a solution to detect and stop cyberbullying [9]. Their solution focuses on detecting and evaluating cyberbullying in the displayed Arabic content. A combination of techniques was used for preprocessing, such as the TweetToSentiSt StrengthFeatureVector filter, converting strings to word vectors, and normalization. Next, the proposed system was trained by Naïve Bayes and SVM models.
Schmidt and Wiegand provided a brief, comprehensive, and structured overview of automatic hate speech detection [10]. Existing methodologies have been systematically identified, with a particular focus on feature extraction. The research is primarily aimed at Natural Language Processing (NLP) researchers who are new to the field of hate speech detection and want to keep themselves up to date with the latest technology. Hee et al. in 2018 concentrated on the automatic recognition of cyberbullying signs that, when employed in a real-life moderation session, require further analysis by human mediators [11]. Binary classification studies were conducted to automatically detect cyberbullying using a linear kernel SVM developed in LIBLINEAR using a machine learning framework for Python and a Scikit-Learn.

Bu et al. proposed an assembly method of two-deep learning models [12]. The first is a character-level convolutional neural network (CNN), which is noise-resistant and extracts low-level syntactic information from a series of characters. The second one uses word-level long-range recurrent convolutional networks (LRCN), which extract high-level semantic information from word sequences. According to experimental data, the performance of the ensembling approaches is greatly improved in identifying cyberbullying comments.

Al-Garadi et al. reviewed the existing literature for detecting aggressive behavior on social media sites using algorithms of machine learning such as (support-vector machines (SVM), k-nearest neighbors (KNN), NB, Random Forest (RF), Decision Tree(DT), LR(L stands for left-to-right scanning of the input stream; R stands for the construction of right-most derivation in reverse), Association Rule Mining (ARM) and red-black algorithm (RB)) [13]. They review four aspects of detecting cyberbullying messages by using the approaches of machine learning: collecting data, engineering the features, building a cyberbullying detection model, and evaluating cyberbullying detection models.

Hani et al. suggested an approach for cyberbullying detection, they evaluated their model using two classifiers, SVM and Neural Network (NN), after using TF-IDF and sentiment analysis algorithms for feature extraction [14]. Tokenization, text reduction, stop word removal, and word collection were some of the preprocessing operations carried using the Microsoft Bing word correction API. The categorizations were assessed using several n-gram language models. While employing TF-IDF and sentiment analysis, they achieved 92.8% accuracy using NN with 3-grams and 90.3% accuracy using SVM with 4-grams. In addition, they found that their NN conducted better than the SVM classifier as it also attained an average f-score of 91.9%, while the SVM achieved an average f-score of 89.8% [15], suggested a personal framework called PI-Bully for identifying cyberbullying. It is based on psychological research that emphasizes the traits of both bullies and victims that make them stand out, as well as the influence of peers who share similar interests in identifying cyberbullying behaviors. They compared PI-Bully with common text classification models such as (KNN, Random Forest, Linear SVM, and Logistic Regression) with the same input features and two text-based cyberbullying detection models, and outperformed them. [16] Suggest the use of sentiment analysis and lexical techniques for the automatic detection of cyberbullying. The Java programming language was used to construct this experimental project. Datasets were gathered via the Twitter API, Microsoft Flow, and YouTube comments. After cleaning and pre-processing the data, they categorized the data as either bullying or non-bullying. They then combined the collected dataset into a single file that had around 100,327 tweets and comments. Three people classified it, and the final classification—which came after the majority classification—was done by an odd number of persons. They employed Pointwise Mutual Information (PMI), Chi-square, and Entropy after the data was finished and set up for use in lexicon development. In comparison to Chi-square and Entropy techniques, the results demonstrate that the PMI approach performs best at detecting cyberbullying.

Beckman et al. described a systematic review that looked at how cyberbullying affected kids and teens who needed special schooling because of psychiatric or psycho-intellectual issues [17]. Cyberbullying may be practiced by students with impairments brought on by intellectual disabilities (IND). According to the studies that were cited, cyberbullying is spread by students with IND at rates between 0% and 41%, cybercrime is committed at rates between 0% and 16.7%, and bullying victims makeup 6.7% of the total. Each study used a cross-sectional design, with a sample size of 22 to 149 participants aged 10 to 21 who had an IND. The impairment diagnostic fell under the umbrella of mental and neuropsychiatric illnesses.

Joshi et al. in 2020 tried to find an optimal algorithm to detect cyberbullying [18]. Twitter was selected as a dataset because it generates many data every day. They change all the uppercase words to lowercase and remove all the punctuation marks and emojis in the preprocessing step, followed by the word embedding technique as a feature extractor. Two methods are used, (SVM) and (CNN). CNN accuracy yields more precise output than SVM. [19], The protective factors and perilous agents for cyberbullying were investigated. One goal was to study the relevance between peer relationships, negative feeling regulation strategies, and cyberbullying. The second objective is to inspect the causative relationship between cyberbullying and cyberbullying. The study was descriptive research using cross-sectional and linear data. Data were analyzed using structural equation modeling (SEM) and hierarchical regression analysis. Study findings have implications for preventing cyberbullying.

Patel et al. adopted data mining techniques of the concerned survey results and converts them into knowledge using a self-report questionnaire [20]. Different potential preprocessing steps could be done by removing duplicate IDs, removing special characters, and tokenizing. Several ML techniques are used to classify the bullying comments, like K-Nearest-Neighbor, SVM, Random Forest Regression, and Logistic Regression. Nirmal et al. pointed out to the main aim of detecting the cyberbullying model that will help improve manual monitoring of online bullying on social networks [21].

Fortunatus et al. collected the dataset of bullying words, fetch the tweets from Twitter, preprocess it, apply natural language processing, and then use SVM as a machine learning classifier to detect the cyberbullied tweets from normal ones [22]; They chose their data from Melania Trump's Facebook comments; they extracted data using scraping software written in Python using the Facebook Graph API. Two hundred and two random entries were selected as a training data set. These entries were annotated manually, resulting in 61 non-aggressive and 141 aggressive entries. Another 403 entries were chosen randomly and annotated manually by two native English speakers as a test data set. The result of manual annotation was then evaluated using performance measures of accuracy, accuracy, recall, and F1 score.

Sultan et al. in 2021 introduced this article to examine articles on cyberbullying of kids and young people and lay out an approach for activity, primarily connected with Kazakh online entertainment [23]. Also, the article gives instances of comparable work by unfamiliar specialists in England, the United States, and other European nations. Utilizing hypothetical techniques, that is to say, investigation, combination, and exact: correlation and examination, chips away at this theme were dissected and determined to break down the issue in the Kazakh space. So, they take a gander at the instance of making an analyzer and gathering information to prepare AI and profound learning calculations that can recognize and impede scripts with a paresthesia incline continuously. The review planned to make an analyzer of the Kazakh language and to concentrate on the pervasiveness of Cyberbullying among youngsters and teenagers with insane or psycho-scholarly issues who need a specialized curriculum. They incorporated a set (a pack of articulations) of cyberbullying words in the Kazakh language, handled and tidied up fundamental data, and fixed it for the errand of organizing equal substance. Cyberbullying is harassment utilizing dreary advanced innovations, which are planned to threaten, outrage, or uncover the objective. Feeling helpless, defenseless, embarrassed, detached, discouraged, or self-destructive are instances of the pessimistic impacts of cyberbullying.

Phanomtip et al. showed that one-third of the understudies have encountered cyberbullying in the course of their life, so they propose a dataset of 67K tweets gathered from Twitter, TF-IDF, and doc2vec are utilized to remove dataset highlights, linear SVM was utilized to recognize the toxic behaviors [24]. The exploratory outcomes exhibit that their technique beats the baseline by 4%.
Gada et al. utilized word2vec to prepare the custom word, whereupon they consolidated two deep learning designs, LSTM (long-transient memory organization) and CNN [25]. Their LSTM - CNN architecture. They test their model on Twitter posts and remarks. Additionally, they created a web application that used the model to categorize tweets as cyberbullying or not based on the poisonousness score as well as other features. The approach was also applied to the Telegram Bot, which monitors and prevents cyberbullying. To remove and redact Not Safe For Work (NSFW) content and stop cyberbullying on WhatsApp Web, two Chrome Extensions were developed. They achieved an amazing result with a 97% ROC AUC score for their model.
Ali et al. compared the predictive models in cyberbullying disclosure between the fundamental machine learning system and their proposed framework with the inclusion of component choice procedure, resampling, and hyperparameter enhancement by using two classifiers: Support Vector Classification Linear and Decision Tree [26]. Before being used in various examination arrangements, word n-gram characteristics from the ASKfm corpus were elicited. The Decision Tree performs best when using highlight selection without resampling and hyperparameter advancement association, according to an analysis of performance metrics.

Emon et al. in 2022 suggested a model for locating Bengali-language cyberbullying on social media [27]. On a dataset of 44,001 Bangla comments from Facebook, they used a variety of transformer models, including Bangla Bidirectional Encoder Representations from Transformers (BERT), Bengali DistilBERT, and Cross-lingual Language Model (XLM-RoBERTa). Among the models, the XLM-RoBERTa model had the best accuracy rate (85%) and F1 score (86%).

Raj et al. proposed a deep learning framework that will evaluate cyberbullying in real-time Twitter tweets [28]. Three datasets were collected from different resources containing English texts, Hindi texts, and the last one contains a combination of Hindi and English texts. The CNN-BiLSTM network has the best accuracy with 95%. Saadi et al. in 2023 introduced a model for detecting cyberbullying on Twitter for Arabic language [29]. Two classifiers were used, Extreme Gradient Boosting (XGBoost) and RF. Cuckoo search optimizers were used to improve their accuracy, with 0.867 for both methods. In [30], negative Arabic comments were detected using the SVM algorithm. For feature extraction, the TF-IDF and the count vectorizer methods were used and then improved using the cuckoo search algorithm. 85.8% and 87.1% are the resulting accuracy before and after optimizing the SVM's hyperparameters, respectively. Sangeethapriya et al. identified the bullying comments in text, image, and video form [31]. They recognize images using a graph convolutional neural network (GCN) and a pre-trained Googlenet.they also use a Mel-scale filter bank speech spectrogram and CNN network model for audio post-classification. This study yielded an accuracy of 96%.

3. Methodology

All the above researches have three steps, theses are preprocessing, feature extraction, and classification. Preprocessing Step: The preprocessing technique reduces unnecessary data and makes it more classifiable. Certainly, there are many pre-processing operations done by different researchers to clean the datasets (special characters, punctuation, white spaces, contractions, and stop words), convert all characters to lowercase, as well as return the words to their root using stemming or lemmatization [32-34]. Feature Extraction: A PC can utilize advanced information. Thus, it is important to decipher the data on the PC as indicated by the language by message portrayal. The message portrayal process is one of the significant in regular language handling examinations utilizing strategies like TF-IDF [4, 14, 15, 21, 24, 35, 36], and weighted TF-IDF [5]. Sentences are represented as dense word vectors via word embedding, hence the term "embedding" refers to obtaining more data with fewer dimensions. [32]. The embedding process translates semantic meaning into geometric meaning, Word2Vec [8, 23, 25], Global vectors (GloVe) for word representation [18], n-grams [11, 26], word-embedding [12], BOW [7] are the pioneers of word representativeness approaches.
Classification Step: Different methods of machine learning and deep learning are used by different researchers, and machine learning algorithms, such as SVM [3, 4, 7, 9, 11, 14, 15, 18, 21, 24, 26, 37-39], Decision tree [1, 4, 13, 26], logistic regression [15, 20, 25], Naïve Bayes [9, 40], random forest [25, 41, 42], XGBoost [25]) were applied. Recently, deep learning methods are used like CNN [12, 18, 43], a hybrid of CNN-LSTM [25], and LRCN [12].

4. Publication Analysis

Published cyberbullying articles increments progressively in the last ten years. Thus, in this part, we attempt to contrast the published research articles on cyberbullying from the principal distributed article in Science Direct from 2011 until 2022. Graphs are utilized to analyze the published articles considering the examination articles being the larger part in their numbers according to the charts, as displayed in Figure 1.

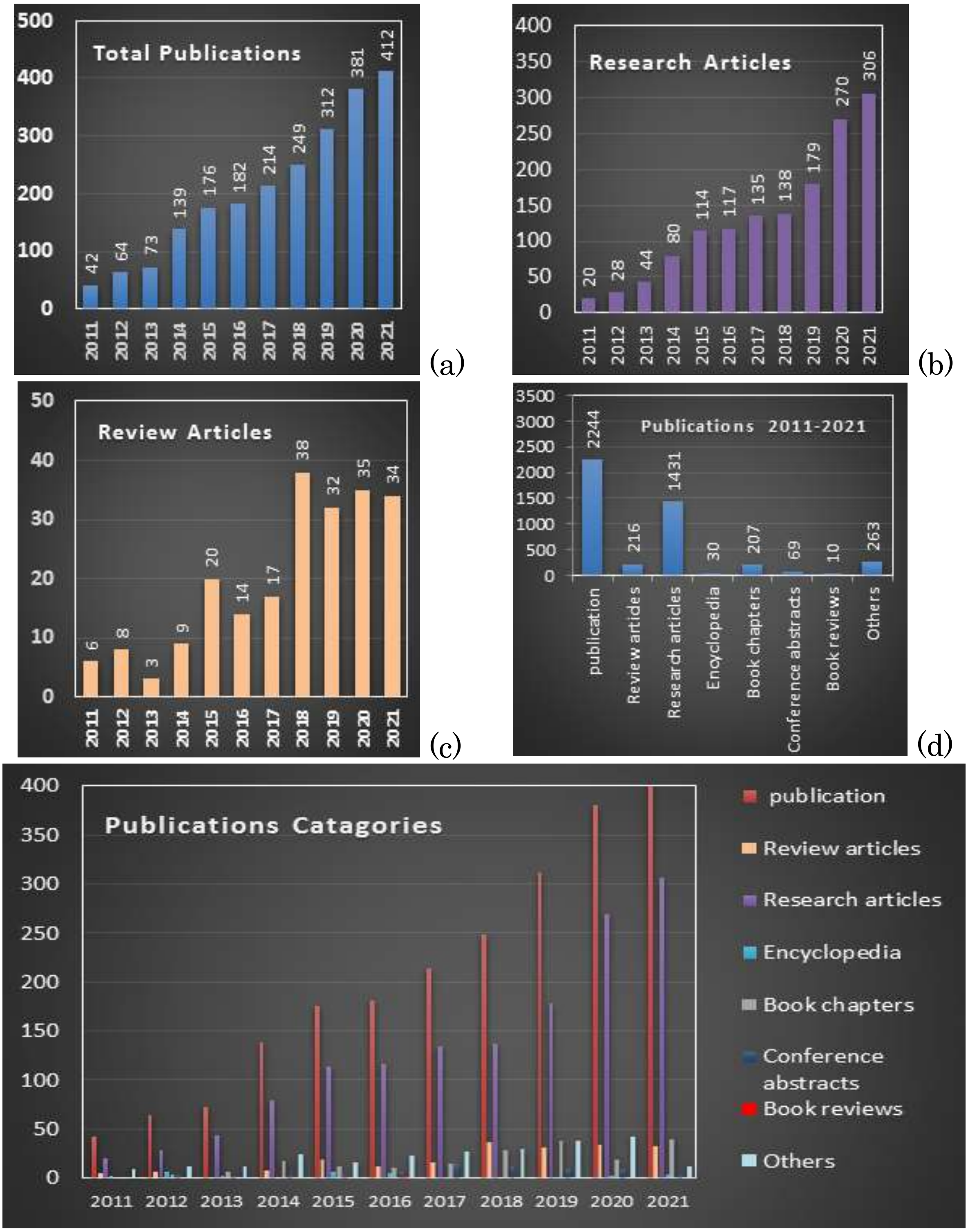
5. Analysis of the Published Research

An analysis is introduced in table form for each paper, in terms of the analysis, publication year, authors, data set, data set size, and approach, as the precision of the outcomes is introduced in Table no 1 which shows twenty published papers concentrated on cyberbullying where they applied different strategies
strategies on various data sets. Accuracy measures were utilized in a large portion of the outcomes, and other metrics are mentioned when used. From Table 1, the process of cyberbullying detection in general consists of well-known steps. These are reading datasets, preprocessing steps, feature extraction, classification, and finally, evaluation.

6. Datasets Collection

The cyberbullying datasets are obtained in different ways. The researchers chose other platforms for their research, such as:

- YouTube is the world's largest user-driven video content provider [38, 44, 45].
- Twitter is a social media tool that lets people share information, in a real-time news feed, with like-minded persons [46-50].
- Netflix is an international streaming video service with over 125 million subscribers in over 190 countries that can be accessed on computers, cell phones, tablets, and televisions [51]. And other platforms.
  With the impediment of after-school and summer exercises because of social removal limitations, numerous children are going to cell phones, tablets, and different gadgets to keep them engaged.



(e)

Figure 1. A summary of published works on cyberbullying. (a)Total publications, (b) Research Articles, (c) Review Articles, (d) Publications 2007-2021, (e) Different publication categories.

Table 1:Summary of Literatures Review

<table>
<tr><th>Year</th><th>Ref.</th><th>Dataset</th><th>Size</th><th>Extracted features</th><th>Method / Technique</th><th>Metrics</th></tr>
<tr><td>2011</td><td>[1]</td><td>Formspring.me</td><td>3915 posts</td><td>• number of “bad” words
• density of “bad” words</td><td>C4.5 decision tree</td><td>Acc. = 78.5%</td></tr>
<tr><td rowspan="4">2012</td><td rowspan="4">[2]</td><td>MySpace</td><td rowspan="4"></td><td rowspan="4">• bag-of-word
• PLSA</td><td rowspan="4">Linear SVM</td><td>Acc = 97.97%</td></tr>
<tr><td>SlashDot</td><td>Acc = 99.49%</td></tr>
<tr><td>Kongregate</td><td>Acc = 100%</td></tr>
<tr><td>Combined datasets</td><td>Acc = 99.20%</td></tr>
<tr><td>2012</td><td>[3]</td><td>Twitter posts</td><td>3,001,427 emotions in 7-classes</td><td>Unigrams, and bigrams tweets</td><td>SVM</td><td></td></tr>
<tr><td>2014</td><td>[4]</td><td>Youtube videos</td><td>54,050 comments</td><td>• Content features
• TF-IDF</td><td><table><tr><th>Approach</th><th>Hybrid app</th></tr><tr><td>SVM
NB
DT
MCES</td><td>+ MCES (H1)
+ MCES (H1)
+ MCES (H1)
+ Naïve Bayes (H2)</td></tr></table></td><td><table><tr><th>AUC for Approach</th><th>AUC for Hybrid app</th></tr><tr><td>59%
66%
52%
72%</td><td>68%
72%
54%
76%</td></tr></table></td></tr>
<tr><td>2015</td><td>[5]</td><td>Twitter Filter API</td><td>30000 tweets</td><td>weighted TF-IDF</td><td>LDA+SVM</td><td>recall = 85.54%</td></tr>
<tr><td rowspan="2">2015</td><td rowspan="2">[6]</td><td>MySpace</td><td rowspan="2"></td><td rowspan="2">Fuzzify rule set, then use GA optimizer</td><td rowspan="2">Naïve classifier</td><td>Acc = 87%</td></tr>
<tr><td>Formspring.me</td><td>Acc. = 86%</td></tr>
<tr><td rowspan="3">2016</td><td rowspan="3">[7]</td><td>Formspring.me</td><td>20.921 questions and answers</td><td rowspan="3">BoW</td><td rowspan="3">GHSOM</td><td>Acc = 73%</td></tr>
<tr><td>YouTube</td><td>3.045 posts</td><td>Acc = 69%</td></tr>
<tr><td>Twitter</td><td>1000 tweets</td><td>Acc = 72%</td></tr>
<tr><td>2016</td><td>[8]</td><td>Instagram</td><td>3000 English comments</td><td>LDA, Offensiveness score, BoW, Word2Vec</td><td>SVM + RBF kernel</td><td>Acc = 95%</td></tr>
<tr><td rowspan="2">2017</td><td rowspan="2">[9]</td><td rowspan="2">Tweets post</td><td rowspan="2">35273 Arabic Tweets</td><td rowspan="2"></td><td>Naïve Bayes</td><td>recall = 90.9%</td></tr>
<tr><td>SVM</td><td>Recall = 92.7 %</td></tr>
</table>

2018	[11]	ASKfm	113,698 English corpus	• word n-grams • subjectivity lexicons • character n-grams • term lists • topic models	Linear SVM classifier	F-score= • 60.09% • 56.82% • 52.69% • 40.48% • 17.35%
			78,387 Dutch corpus			F-score= • 55.52% • 54.34% • 51.70% • 28.65% • 24.74%
2018	[12]	The pseudo-badword	8,815 comments	• word-embedding	• CNN • LRCN • Hybrid	AUC • 85.41% • 85.73% • 87.22%
2019	[14]	Cyberbullying dataset from Kaggle	12773 messages	• TF-IDF • Sentiment Analysis	• SVM • Neural Network	Acc = 90.3% Acc= 92.8%
2019	[15]	microblogging platform,	3095 posts	• Linguistic Inquiry Word Count • TF-IDF	• kNN • RF • SVM • LR • Pi-Bully	AUC • 65.2% • 70.1% • 70.7% • 70.5% • 84.4%
		Twitter	19,994 labeled tweets			• 66.2% • 71.4% • 67.8% • 71.1% • 80%
2019	[16]	Twitter API, Microsoft-Flow, and YouTube comments	100,327 Arabic tweets and comments		• PMI • Chi-square, • Entropy	F-score • 81% • 62.11% • 39.14%
2020	[18]	Twitter		• Glove embedding	• SVM Unigram • SVM Bigram • SVM Trigram • SVM n-gram • CNN	Acc= 84% Acc= 86.9% Acc= 85.1% Acc= 87% Acc= 90.2%
2020	[21]	Twitter	1762 tweets	• Bigram TF-IDF	SVM	Acc= 81.3 %
2020	[22]	Melania Trump's Facebook post comments	403 entries	• POS	lexicon enhanced rule-based	Acc= 71.782%

2021	[24]	Twitter	67K tweets	• TF-IDF • Doc2Vec.	Linear SVM	Acc= 91% Acc= 86%
2021	[25]	Wikipedia comments	1.6 million rows	• word2vec	• Logistic Regression • Random Forest • XGBoost • LSTM - CNN	Acc= 92% Acc= 92% Acc= 91.6% Acc= 95.2%
2021	[26]	AMiCA Bullying Cyber dataset from ASKfm	106,477	Word n-grams bag-of-words (BoW)	SVM (SVC Linear)	Acc= 95.74 %
					DT	Acc= 98.13%
2022	[27]	Bangla Text Dataset	44,001	• BERT	XML-RoBERTa	Acc= 85%
2022	[28]	Twitter from Formspring and Wikipedia on English, Hindi languages		• GloVe and FastText	CNN-BiLSTM	Acc= 95%
2023	[28]	Twitter for Arabic language	7,748	• TF-IDF	XGBoost	Acc= 86.7%
					RF	Acc= 86.7%
2023	[29]	Twitter for Arabic language	7,748	• TF-IDF and count vectorizer	SVM	Acc= 87.1%
2023	[30]	no datasets were generated or analyzed during the current study			GCN and Mel-frequency	Acc= 96%

New popular assessment information from Morning Consult investigates what children are doing on the web and which stages are encountering a leap in extra screen time. As shown in Figure 2, Among all age groups, the two most well-known entertainment platforms were YouTube and Netflix Inc. 55% of parents stated something similar about Netflix. In contrast, 62% of parents said their kids interacted with YouTube. However, activity on YouTube Kids, the video platform's kid-focused website, slowed, with only 23% of parents generally reporting that their kids used the platform. Virtual entertainment stages were more well-known among teenagers. Among guardians of teens, 40% said their kids utilized Instagram, while 36% said something similar to TikTok.

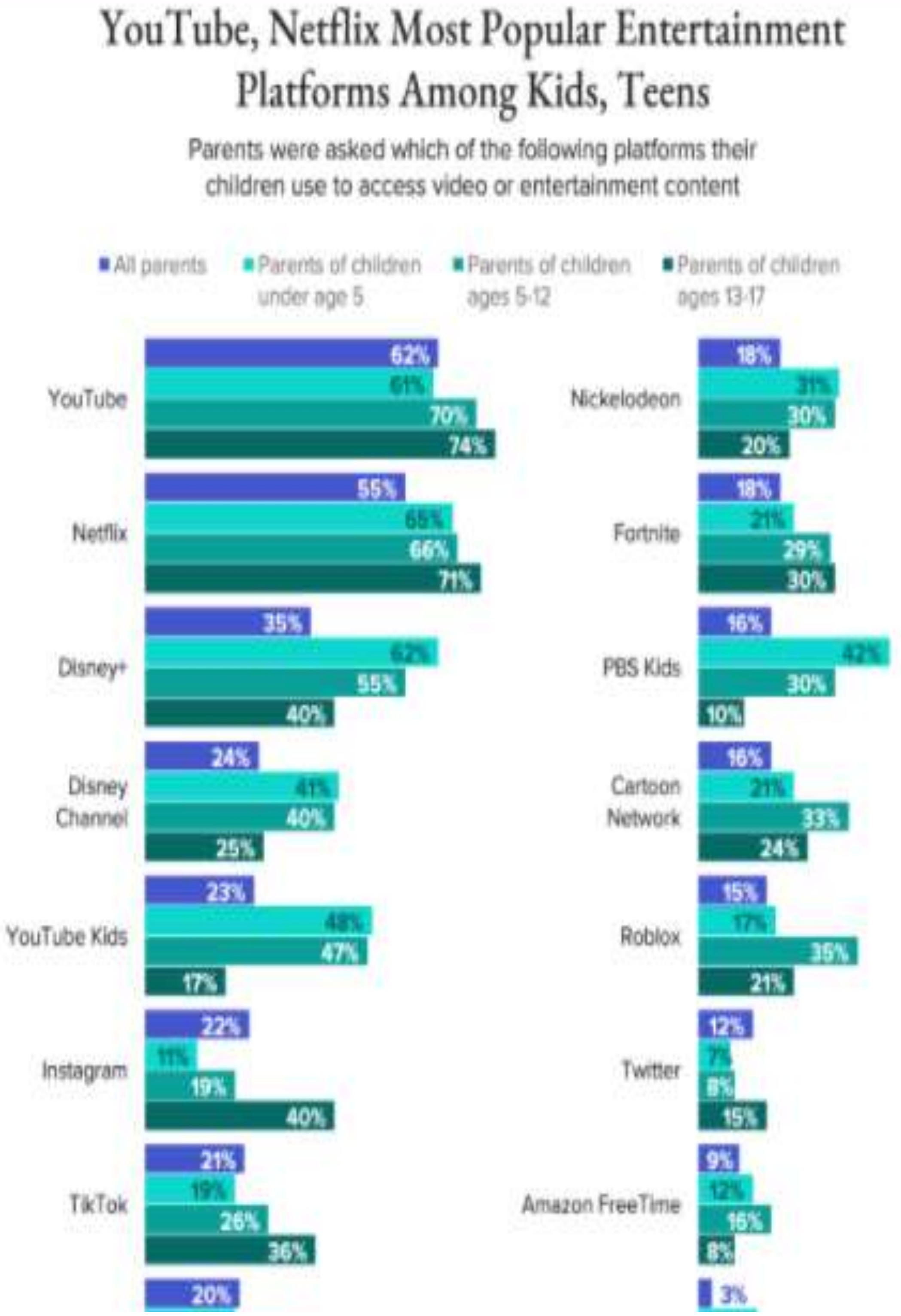


Figure 2. Poll Conducted Aug. 13-15, 2020, Among 899 parents whose child has access to an electronic device, with a margin of error of +/-3%. According to Morning Consult [55]

- Kongregate (An online gaming website and/or social network), this gaming site besides games, has other interactive features that capable users to create a profile, chat while gaming, and post in forums, they assume the players while playing use aggressive words [53].
- Facebook is an interactive network where users can share their thoughts and information over a wide area [52].
- A social networking site called MySpace was established in 2003. This website has always been heavily reliant on music. This website used to be the most well-known of its sort [54].
- Formspring: This is a question-answer-based online platform where an anonymous person can also comment [49, 55].

7. Discussion

The main aim of this research is to examine the methods for identifying cyberbullying, which has unquestionably become more prevalent in recent years. With various metrics, various word embedding, machine learning, and deep learning algorithms were applied. Many datasets are found dealing with cyberbullying, but most of them are unbalanced English datasets. The researchers used different rating scales depending on their goals. Understanding the nature of manually categorised data before selecting a rating system is vital. Many researchers use accuracy as the primary rating scale. The authors in [5] employed oversampling of minority cases to enhance classifier training. Researchers may need to utilize AUC as the significant evaluation metric because it is more dependable than other performance metrics in circumstances where the data are uneven. Some researchers focus on recall measures [2] since they are willing to increase the percentage of true positives by accepting false positives that the tool identifies, others choose AUC as the main assessment measure, which is more robust than other performance measures. Machine learning methods are outperformed by deep networks in the same data set, as reported by [18] and [25]. The performance of each deep network is optimized through methods of knowledge transfer and word embedding. The research has some limitations as it deals only with articles written in English and published in full text. This type of study provides a starting point for researchers in revealing the main consensus on the topic examined and in the context of presenting an increasing number of publications to researchers.

8. Conclusions

Cyberbullying has increased exponentially since 2018.The increasing phenomenon of cyberbullying among young people and children is due to the increasing need for e-learning among school and university students, especially after the Corona pandemic. Deep learning models are more efficient than machine learning due to the upward inflation of bullying data. Geographical characteristics, such as users' locations, as well as temporal characteristics, such as the times of their activities, may reveal more details about the frequency of bullying occurrences, common places, and bullying-related periods. Due to Twitter's character cap restrictions, users are only able to give the bare minimum of information. Short papers constitute a significant barrier for text analysis because the most popular techniques are most effective for long documents.

9. Recommendation for Future Works

Build an Arabic cyberbullying dataset from YouTube or Twitter, and try to find the cyberbullying comments. Use deep learning models such as Bi-directional Long-Short Term Memory (Bi-LSTM), or Bidirectional Encoder Representations from Transformers (BERT) models. Solve the unbalanced datasets through up-sampling techniques using Synthetic Minority Oversampling Technique (SMOTE), or down-sampling through deleting random comments from the major class.

Acknowledgments: The authors would like to thank the Computer Science Department\ College of Science and Al-Nahrain University for their support.

Funding: This research is self-funded.

Conflict of Interest: The authors declare no conflict of interest.

References

[1] Reynolds, K.; Kontostathis, A.; Edwards, L.; "Using machine learning to detect cyberbullying". Proceedings of 10th International Conference on Machine Learning and Applications and Workshops,

Honolulu, HI, USA, 18-21 December , IEEE; 2011.
[2] Vinita, N.; Sayan, U.; Xue, L.; Pang, A.C.; "Sentiment Analysis for Effective Detection of Cyber Bullying".Springer, 7235: 7, 2012.
[3] Xu, J.-M.; Zhu, X.; Bellmore, A.; "Fast learning for sentiment analysis on bullying". Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, Beijing China, August, Association for Computing Machinery, New York NY, United States; 2012.
[4] Dadvar, M.; Trieschnigg, D.; de-Jong, F.; "Experts and machines against bullies: A hybrid approach to detect cyberbullies". Proceedings of 27th Canadian Conference on Artificial Intelligence, Canadian AI 2014, Montréal, QC, Canada, May 6-9, Springer Cham, 2014.
[5] Nalini, K.; Sheela, L.J.; "Classification of tweets using text classifier to detect cyber bullying". Springer, 2 : 637-645:,2015.
[6] Nandhini, B.S.; Sheeba, J.; "Online social network bullying detection using intelligence techniques". Procedia Comput Sci, 45: 485-492, 2015.
[7] Capua, M.D.; Nardo, E.D.; Petrosino, A.; "Unsupervised cyber bullying detection in social networks". Proceedings of 23rd International Conference on Pattern Recognition (ICPR) Cancún Center, Cancún, México, December 4-8, 2016.
[8] Zhong, H.; Li, H.; Squicciarini A. and Rajtmajer S.; "Content-Driven Detection of Cyberbullying on the Instagram Social Network". Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, USA 9–15 July, AAAI Press / International Joint Conferences on Artificial Intelligence, Palo Alto, California USA, 2016.
[9] Haidar, B.; Chamoun, M.; Serhrouchni, A.; "A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning". Adv. Sci. Technol. Eng. Syst. J., 2(6): 275-284, 2017.
[10] Schmidt, A.; Wiegand, M.; "A survey on hate speech detection using natural language processing". Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Valencia, Spain, April, Association for Computational Linguistics,2017.
[11] Hee, V.C.; Jacobs, G.; Emmery, C.; Desmet, B. Lefever, E.; Verhoeven, B. et al.; "Automatic detection of cyberbullying in social media text". PLoS One, 13(10): e0203794, 2018.
[12] Bu, S.; Cho, S.; "A hybrid deep learning system of CNN and LRCN to detect cyberbullying from SNS comments". Proceedings of 13th International Conference, HAIS, Oviedo, Spain, June 20-22, Oviedo, Spain, 2018.
[13] Al-Garadi, M.A.; Hussain, M.R.; Khan, N.; Murtaza G.; "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.
[14] Hani, J.; Nashaat, M.; Ahmed, M.; Emad, Z.; Amer, E.; Mohammed, A.; "Social media cyberbullying detection using machine learning". Int. J. Adv. Comput. Sci. Appl, 10(5): 703-707, 2019.
[15] Cheng, L.; Li, J.; Silva, Y.N.; Hall, D.L.; Liu, H.; "PI-Bully: Personalized Cyberbullying Detection with Peer Influence". Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence,Macao, 10-16 August , International Joint Conferences on Artificial Intelligence,2019.
[16] Al-Harbi, B.Y.; Al-Harbi, M.S.; Al-Zahrani, N.J.; Alsheail, M.M.; Alshobaili, J.F.; Ibrahim, D.M.; "Automatic cyberbullying detection in Arabic social media". Int. J. Eng. Res. Sci. Technol., 12(12): 2330-2335, 2019.
[17] Beckman, L.; Hellström, L.; von-Kobyletzki, L.; "Cyberbullying among children with neurodevelopmental disorders: A systematic review". Scand. J. Psychol., 61(1): 54-67, 2020.
[18] Joshi, P.I.R.; Kaulgud, N.; Suryawanshi, A.; Lokhande, M.; "Detecting cyber bullying on Twitter using machine learning techniques". Proceedings of 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 13-15 May, IEEE, 2020.
[19] Akgül, G.; Artar, M.; "Cyberbullying: relationship with developmental variables and cyber victimization". Scand. J. Child. Adolesc Psychiatr Psychol, 8(4): 25-38, 2020.
[20] Patel, M.; Sharma, P.; Cherian, A.K.; "Bully identification with machine learning algorithms". J. Crit. Rev., 7(6): 417-425, 2020.
[21] Nirmal, N.; Sable, P.; Patil, P.; Kuchiwale, P.S.; "Automated Detection of Cyberbullying

Using Machine Learning". Int. Res. J. Eng. Technol.,, 07(12): 2054- 2061, 2020.
[22] Fortunatus, M.; Anthony, P.; Charters, S.; "Combining textual features to detect cyberbullying in social media posts". Procedia Comput. Sci.,176: 612-621, 2020.
[23] Sultan, D.; Suliman, A.; Toktarova, A.; Omarov, B.; Mamikov, S.; Beissenova, G.; "Cyberbullying Detection and Prevention: Data Mining in Social Media", Proceedings of 11th International Conference on Cloud Computing, Data Science & Engineering, Noida, India, 28-29 January, IEEE, 2021.
[24] Phanomtip, A.; Sueb-in, T.; Vittayakorn, S.; "Cyberbullying detection on Tweets". Proceedings of 18th International Conference onElectrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Chiang Mai, Thailand,19-22 May, IEEE, 2021.
[25] Gada, M.; Damania, K.; Sankhe, S.; "Cyberbullying Detection using LSTM-CNN architecture and its applications". Proceedings of 2021 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 27-29 January, IEEE, 2021.
[26] Ali, W.N.; Mohd, M.; Fauzi, F.; "Cyberbullying Predictive Model: Implementation of Machine Learning Approach". Preceedings in fifth International Conference on Information Retrieval and Knowledge Management (CAMP), Kuala Lumpur, Malaysia, 15-16 June, IEEE, 2021.
[27] Emon, M.; Iqbal, K.N.; Mehedi, M.; Mahbub, M.J.; Rasel, A.A.; "Detection of Bangla Hate Comments and Cyberbullying in Social Media using NLP and Transformer Models". Proceedings of International Conference on Advances in Computing and Data Sciences, Kumool, India, 22-23 April, Springer, 2022.
[28] Raj, M.; Singh, S.; Solanki, K.; Selvanambi, R.; "An Application to Detect Cyberbullying Using Machine Learning and Deep Learning Techniques". SN comput. sci., 3(401): 1-13,2022.
[29] Saadi, M.Q.; Dhannoon, B.N.; "Comparing the Random Forest vs. Extreme Gradient Boosting using Cuckoo Search Optimizer for Detecting Arabic Cyberbullying". Iraqi J. Sci. 64(9): 4806-4818, 2023.
[30] Saadi, M.Q.; Dhannoon, B.N.; "Arabic Cyberbullying Detection Using Support Vector Machine with Cuckoo Search". Iraqi J. Sci., 64(10): 5322- 5330, 2023.
[31] Sangeethapriya, R.; Akilandeswari, J.; "Classification of cyberbullying messages using text, image and audio in social networks: a deep learning approach". Multimed Tools Appl.,38:2237–2266, 2024.
[32] Hamzah, N.A.; Dhannoon, B.N.; "The Detection of Sexual Harassment and Chat Predators Using Artificial Neural Network". Karba. Int. J.Mod. Sci., 7(4): 301-312, 2021.
[33] Aldhyani, T.H.; Al-Adhaileh, M.H.; Alsubari S.N.J.E.; "Cyberbullying identification system based deep learning algorithms". Electronics,11: 1-19, 2022.
[34] Iwendi, C.; Srivastava, G.; Khan, S.; Maddikunta, P.K.; "Cyberbullying detection solutions based on deep learning architectures". Multimed. Syst. 29(9) :1839–1852, 2023.
[35] Alhashmi, A.; Darem, A.B.; "Consensus-Based Ensemble Model for Arabic Cyberbullying Detection". Comput. Syst. Sci. Eng.,41(1): 241-254, 2022.
[36] Rachid, B.A.; Azza, H.; Ghezala, H.H.B.; "Classification of cyberbullying text in Arabic", Proceedings of International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19-24 July, IEEE,2020.
[37] Guerrero, M.D.; Barnes, J.D.; Chaput, J.-P.; "Tremblay, M. S. Comparison of Machine Learning Techniques for Cyberbullying Detection on YouTube Arabic Comments" . Int. J. Comput. Sci. Network Secur., 21 (1), 1–5, 2021.
[38] Michelle, D.G.; Joel, D.B.; Jean-Philippe, C.; Mark, S.T.; Michelle, D.G.; Joel, D.B.; Jean-Philippe, C.; Mark, S.T.; Almutiry, S.; Abdel Fattah, M.J.T.; "Arabic CyberBullying Detection Using Arabic Sentiment Analysis". Egypt. Inform. J., 8(1): 39-50, 2021.
[39] AL-Bakri, N.F.; Al-Zubidi A.F.; Alnajjar, A.B.; Qahtan, E.; "Multi label restaurant classification using support vector machine". Period. Eng. Nat. Sci., 9(2): 774-783, 2021.
[40] Mouheb, D.; Albarghash, R.; Mowakeh, M.F.; Al-Aghbari, Z.; Kamel, I.; "Detection of Arabic cyberbullying on social networks using machine learning". Proceedings of IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), Abu Dhabi, United Arab Emirates, 03-07 November,IEEE,2019.

[41] Ali, A.; Syed, A.M.; "Cyberbullying detection using machine learning". Pak. J. Eng. Appl. Sci., 3(2): 45-50, 2020.

[42] Wotaifi, T.A.; Dhannoon, B.N.; "Improving Prediction of Arabic Fake News Using Fuzzy Logic and Modified Random Forest Model". Karba. Int. J. Mod. Sci., 8(3): 477-485, 2022.

[43] Banerjee, V.; Telavane, J.; Gaikwad, P.; Vartak, P.; "Detection of cyberbullying using deep neural network". Proceedings of 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 15-16 March, IEEE, 2019.

[44] Wattenhofer, M.; Wattenhofer, R.; Zhu, Z.;"The YouTube social network". Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, Trinity College in Dublin, Ireland, June 4–7,2012.

[45] Alsubait, T.; Alfageh, D.; "Comparison of Machine Learning Techniques for Cyberbullying Detection on YouTube Arabic Comments". Int. J. Comput. Sci. Network Secur.; 21(1): 1-5, 2021.

[46] Br J. N.; "Using Twitter as a teaching tool". Training M. V. C., 20(20): 292-6, 2011.

[47] Al-Ajlan, M.A.; Ykhlef, M.; "Optimized Twitter cyberbullying detection based on deep learning"; Proceedings of 21st Saudi Computer Society National Computer Conference (NCC), Riyadh, Saudi Arabia, 25-26 April ,IEEE,2018.

[48] Hadiya, E.; "Cyber Bullying Detection in Twitter using Machine Learning Algorithms". Int. J. Adv. Eng. Manag., 4(8): 1172-1184, 2022.

[49] Mahat, M.; "Detecting cyberbullying across multiple social media platforms using deep learning". Proceedings of International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 04-05 March, IEEE,2021.

[50] Haidar, B.; Chamoun, M.; Serhrouchni, A.; "Arabic cyberbullying detection: Using deep learning". Proceedings of 7th International Conference on Computer and Communication Engineering (ICCCE), Kuala Lumpur, Malaysia, 19-20 September, IEEE, 2018.

[51] Schmidt, G.; "Netflix" In book: The SAGE International Encyclopedia of Mass Media and Society " , Merskin D. L., SAGE Publications, Inc.: Thousand Oaks, 1204-1205, 2019.

[52] https://www.facebook.com/?locale=ar_AR.

[53] https://www.commonsensemedia.org/website-reviews/kongregate.

[54] https://myspace.com

[55] Yadav, J.; Kumar, D.; Chauhan, D.; "Cyberbullying detection using pre-trained bert model". Proceedings of International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 02-04 July,IEEE,2020.