

Article

A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter

Amgad Muneer ^{1,*}  and Suliman Mohamed Fati ² 

¹ Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Seri Iskandar 32160, Malaysia

² Information Systems Department, College of Computer and Information Sciences, Prince Sultan University, Riyadh 11586, Saudi Arabia; smfati@yahoo.com

* Correspondence: muneeramgad@gmail.com

Received: 9 October 2020; Accepted: 20 October 2020; Published: 29 October 2020



Abstract: The advent of social media, particularly Twitter, raises many issues due to a misunderstanding regarding the concept of freedom of speech. One of these issues is cyberbullying, which is a critical global issue that affects both individual victims and societies. Many attempts have been introduced in the literature to intervene in, prevent, or mitigate cyberbullying; however, because these attempts rely on the victims' interactions, they are practical. Therefore, detection of cyberbullying without the involvement of the victims is necessary. In this study, we attempted to explore this issue by compiling a global dataset of 37,373 unique tweets from Twitter. Moreover, seven machine learning classifiers were used, namely, Logistic Regression (LR), Light Gradient Boosting Machine (LGBM), Stochastic Gradient Descent (SGD), Random Forest (RF), AdaBoost (ADB), Naive Bayes (NB), and Support Vector Machine (SVM). Each of these algorithms was evaluated using accuracy, precision, recall, and F1 score as the performance metrics to determine the classifiers' recognition rates applied to the global dataset. The experimental results show the superiority of LR, which achieved a median accuracy of around 90.57%. Among the classifiers, logistic regression achieved the best F1 score (0.928), SGD achieved the best precision (0.968), and SVM achieved the best recall (1.00).

Keywords: cyberbullying detection; tweets classification; Twitter; logistic regression; random forest; light GBM; SGD; AdaBoost; naive bayes; SVM

1. Introduction

Due to the significant development of Internet 2.0 technology, social media sites such as Twitter and Facebook have become popular and play a significant role in transforming human life [1,2]. In particular, social media networks have incorporated daily activities, such as education, business, entertainment, and e-government, into human life. According to [3], social networking impacts are projected to exceed 3.02 billion active social media users each month globally by 2021. This number will account for approximately one-third of the Earth's population. Moreover, among the numerous existing social networks, Twitter is a critical platform and a vital data source for researchers. Twitter is a popular public microblogging network operating in real-time, in which news often appears before it appears in official sources. Characterized by its short message limit (now 280 characters) and unfiltered feed, Twitter use has rapidly increased, with an average of 500 million tweets posted daily, particularly during events [3]. Currently, social media is an integral element of daily life. Undoubtedly, however, young people's usage of technology, including social media, may expose them to many behavioral and psychological risks. One of these risks is cyberbullying, which is an influential social attack occurring on social media platforms. In addition, cyberbullying has been associated with adverse

mental health effects, including depression, anxiety, and other types of self-harm, suicidal thoughts, attempted suicide, and social and emotional difficulties [4–7].

Furthermore, the substantial increase in the number of cyberbullying cases has highlighted the danger of cyberbullying, particularly among children and adolescents, who can be inconsiderate and juvenile. Children and adolescents take bullying seriously without understanding how to manage social issues; this leads them to express their emotions on social media in a manner that can hurt others. According to [8], several studies have shown that bullies often suffer from psychological conditions, leading them to bully and inflict suffering on others. Thus, cyberbullying is similar to an epidemic, and can lead to an aggressive society, particularly regarding high-tech university and school students.

Therefore, many global initiatives have been proposed to tackle the issue of cyberbullying. These initiatives aim to enhance the safety of Internet users, particularly children; for example, the University of Turku, Finland, established an anti-cyberbullying program called Kiva [9], an anti-harassment campaign took place in France [10], and an anti-cyberbully initiative was established by the Belgian government [11]. However, because the Internet's content is vast and difficult to control, detecting and filtering cyberbullies is considered complementary to the legalization and intervention approaches. Thus, detecting cyberbullying in social media is necessary and should be paid high attention so that children and society are protected from its side-effects. Cyberbullying is now a research topic, with researchers aiming to detect, control, and reduce cyberbullying in social media. One direction in this field is to detect a user's intention to post offensive content by analyzing offensive language based on different features, such as the structure and unique content, in addition to the users' writing style. Another direction of cyberbullying research is to detect text content using machine learning for offensive language detection and classification.

Our focus in this comparative study is to review the machine learning classifiers used in the detection of cyberbullying, and examine the performance of these classifiers using accuracy, precision, recall, and F1 score as performance metrics. Therefore, the main contributions of this work are:

- We conducted an extensive review of quality papers to determine the machine learning (ML) methods widely used in the detection of cyberbullying in social media (SM) platforms.
- We evaluated the classifiers investigated in this work, and test their usability and accuracy on a sizeable generic dataset.
- We developed an automated detection model by incorporating feature extraction in the classifiers to enhance the classifiers' efficiency on the sizeable generic dataset.
- We compared the performance of seven ML classifiers that are commonly used in the detection of cyberbullying. We also used the Frequency-Inverse Document Frequency (TF-IDF) and Word2Vec models for feature extraction. This comparison analysis helped to understand the limitations and advantages of ML in text classification models.

Accordingly, we formulated and aimed to answer the following research questions in this work:

- What types of existing machine learning techniques/methods are being used extensively to detect cyberbullying in social media platforms?
- How can an automatic cyberbullying detection model be developed with high accuracy and less processing time?
- How can feature extraction be used to enhance the detection process?

The proposed approach detects cyberbullying by extracting tweets, classifying the tweets using text analysis techniques based on predefined keywords, and then classifying the tweets as offensive or non-offensive. Therefore, the outcomes of the current evaluation will help other researchers to choose a suitable and sufficient classifier for the datasets of global cyberbullying tweets collected from [12,13], because improvements are necessary to further increase the classification accuracy.

This paper is structured as follows. Section 2 is dedicated to the background and related work wherein the examined classifiers will be described. Section 3 provides an overview of the methodology

adopted for the proposed research and a description of the dataset utilized for the experiment. Results are discussed in Section 4, and conclusions and future work are provided in Section 5.

2. Background and Related Work

For several years, the researchers have worked intensively on cyberbully detection to find a way to control or reduce cyberbully in Social Media platforms. Cyber-bullying is troubling, as victims cannot cope with the emotional burden of violent, intimidating, degrading, and hostile messages. To reduce its harmful effects, the cyberbullying phenomenon needs to be studied in terms of detection, prevention, and mitigation.

Presently, there is a range of global initiatives aimed at preventing cyberbully and improving the safety of internet users, including children [14,15]. In the literature, there are many studies to prevent cyberbully in what is called intervention and prevention approaches. Such approaches originate from the psychology and education fields. However, these approaches are globally rare. Besides, cyberbully victims often refuse to speak with a parent [16], teacher [17], or other adults [18]. They spend much time online [19], tend to get anonymous help [20], and post on the Internet a need for information and assistance [21]. However, the effective way of delivering cyberbullying solutions is through the Internet. Web-based approaches can also be used whenever and wherever the patient prefers [22]. For instance, the University of Turku, Finland, has established an anti-cyberbully program called Kiva [9], and Anti-Harassment campaign in France [10], and an anti-cyberbully initiative by the Belgian government [11].

Ideally, these prevention and intervention approaches should: (1) increase awareness of potential cyberbully threats through individualized intensive intervention strategies based on the victims' needs [23–26]; (2) provide health education and teach emotional self-management skills [27]; (3) increase awareness of victims in both reactive measures (e.g., deleting, blocking and ignoring messages), and preventive measures (e.g., increased awareness and security) [28]; provide practical strategies and resources that allow victims to cope with experienced stress and negative emotions [28]; (4) aim to reduce traditional bullying as well [29] since victims are often involved in both forms of bullying [30–32]; and (5) include empathy training, Internet labelling and healthy Internet behavior [33,34]. Thus far, there has been difficulty in preventing cyberbullying. Most parents and teachers rely on the awareness of children on the causes and impacts of cyberbullying. Some parents think that peer-mentoring is an effective way to prevent cyberbullying, particularly in the teenage years, when peers have a more significant impact than the family and school. Therefore, more specific approaches or online resources need to be developed to help the victims [24]. For example, Stauffer et al. [35] provided a prevention caveat stating that bully prevention programs produce a minimal change in student behavior [25].

Similarly, authors in [36] suggest that schools should take the following measures in formulating their cyberbullying prevention program: (1) Define cyberbullying; (2) Have strong policies in place; (3) Train staff, students, and parents on policy identify cyberbullying when they see it; and (4) Use internet filtering technologies to ensure compliance. Past research has indicated that social reinforcement may be a dominant protective factor in mitigating the adverse effects of cyberbullying [37,38]. To get the required reinforcement to minimize the related adverse effects of cyberbullying, they must seek help. However, some reports show that cyberbullying victims are unable to report bullying cases and prefer to be silent [6,39]. Some teenagers rarely seek assistance from their teachers or school advisors [40,41].

Based on the above issues of prevention approaches, the need to detect and filter cyberbullying on social media is highly needed. Thus, this section is dedicated to inspecting cyberbully detection techniques. As per the literature review, there are two main directions in detecting cyberbully: natural Language Processing and Machine Learning, as explained in the following sub-sections.

2.1. Natural Language Processing (NLP) in Cyberbullying Detection

One direction in this field is to detect the offensive content using Natural Language Processing (NLP). The most explanatory method for presenting what happens within a Natural Language Processing system is using the “levels of language” approach [42]. These levels are used by people to extract meaning from text or spoken languages. This levelling refers to the reason that language processing relies mainly on formal models or representations of knowledge related to these levels [42,43]. Moreover, language processing applications distinguish themselves from data processing systems by using the knowledge of the language. The analysis of natural language processing has the following levels:

- Phonology level (knowledge of linguistic sounds)
- Morphology level (knowledge of the meaningful components of words)
- Lexical level (deals with the lexical meaning of words and parts of speech analyses)
- Syntactic level (knowledge of the structural relationships between words)
- Semantic level (knowledge of meaning)
- Discourse level (knowledge about linguistic units more extensive than a single utterance)
- Pragmatic level (knowledge of the relationship of meaning to the goals and intentions of the speaker)

Dinakar et al. [44], for example, used a common-sense knowledge base with associated reasoning techniques. Kontostathis et al. [45] recognized cyberbullying content based on Formspring.me data, using query words used in cyberbullying cases. Xu et al. [46] use several natural language processing methods to detect signs of bullying (a new term relating to online references that could be bullying instances themselves or online references relating to off-line bullying cases). They use sentiment analysis features to identify bullying roles and Latent Dirichlet Analysis to identify subjects/themes. The authors in [46] are intended to set the basis for several tasks relating to identifying bullying and providing a call for other researchers to enhance these specific techniques. Therefore, Yin et al. [47]; Reynolds et al. [48]; and Dinakar et al. [44] are the earliest researchers working in NLP cyberbullying detection, who investigated predictive strength n-grams, part-speech information (e.g., first and second pronoun), and sentiment information based on profanity lexicons for this task (with and without TF-IDF weighting). Similar features were also used for detecting events related to cyberbullying and fine-grained categories of text in [49].

To conclude, some of the common word representation techniques used and proven to improve the classification accuracy [50] are Term Frequency (TF) [51], Term Frequency-Inverse Document Frequency (TF-IDF) [52], Global Vectors for Word Representation (GloVe) [53], and Word2Vec [54]. One of the main limitations of NLP is that of contextual expert knowledge. For instance, many dubious claims about the detection of sarcasm, but how one would detect sarcasm in a short post like “Great game!” responded to a defeat. Therefore, it is not about linguistics; it is about possessing knowledge relevant to the conversation.

2.2. Machine Learning in Cyberbullying Detection

Machine learning-based cyberbullying keywords are another direction of cyberbullying detection, which has been used widely by several researchers. Moreover, Machine learning (ML) is a branch of artificial intelligence technology that gives systems the capability to learn and develop automatically from experience without being specially programmed, often categorized as supervised, semi-supervised or unsupervised algorithms [55]. Several training instances in supervised algorithms are utilized to build a model that generates the desired prediction (i.e., based on annotated/labeled data). In contrast, unsupervised algorithms are not based on data and are mainly utilized for clustering problems [55,56].

Raisi and Huang [57] proposed a model for identifying offensive comments on social networks through filtering or informing those involved. They have used comments with offensive words from

Twitter and Ask.fm to train this model. Other authors [58,59] built communication systems based on smart agents that provide supportive emotional input to victims suffering from cyberbullying. Reynolds [48] suggested a method for detecting cyberbullying in the social network “Formspring,” focused on detecting aggressive trends in user messages, by analyzing offensive words; moreover, it uses a rating level of the threat identified. Similarly, J48 decision trees obtained an accuracy of 81.7%.

Authors in [60] describe an online application implementation for school staff and parents in Japan, with a duty to detect inadequate content on non-official secondary websites. The goal is to report cyberbullying cases to federal authorities; they used SVMs in this work and obtained 79.9% accuracy. Rybníček [61] has proposed a Facebook framework to protect underage users from cyberbullying and sex-teasing. The system seeks to evaluate the content of photographs and videos and the user’s actions to monitor behavioral changes. A list of offensive words was made in [62] using 3915 posted messages monitored from the Formspring.me web site. The accuracy obtained in this study was only 58.5% [62].

Another study [47] suggests a method for identifying and classifying cyberbullying acts as harassment, flaming, terrorism, and racism. The author uses a fuzzy classification rule; therefore, the results are inferior in terms of accuracy (around 40%), but using a set of rules, improved the classifier efficiency by up to 90%.

In [63], authors have developed a cyberbullying detection model based on Sentiment analysis in Hindi-English code-mixed language. The authors carried out their experiments based on Instagram and YouTube platforms. The authors use a hybrid model based on top performers of eight baseline classifiers, which perform better with an accuracy of 80.26% and an f1-score of 82.96%.

Galán-García et al. [64] suggested applying a real case of cyberbullying detection in Twitter using supervised machine learning. The study uses two different feature extraction techniques with various machine learning algorithms, and Sequential Minimal Optimization (SMO) classifier obtained (68.47%), the highest accuracy among the rest. In [65], the authors have proposed a cyberbullying detection approach based on Instagram’s social network. The experiments were carried out based on image contents analysis and user’s comments. The results show that uses multiple features can improve the classification accuracy of linear SVM, where the accuracy of SVM jumped from 0.72 to 0.78 by using image categories as an additional feature. Nahar et al. [66] propose creating a weighted directed graph model for cyberbullying that can be used to calculate each user’s predator and victim scores while using a weighted TF-IDF scheme with textual features (second-person pronouns and foul words) to improve online bullying.

Salminen et al. [67] suggest a hate content detection approach for multiple social media networks. The authors use a total of 197,566 comments from four platforms: YouTube, Reddit, Wikipedia, and Twitter, with 80% of the comments labelled non-hateful, and the remaining 20% was hateful. The experiments were conducted using several machine learning algorithms to test each feature separately to evaluate their accuracy based on features selection. In addition to machine learning classifiers, Dadvar et al. [68] suggested an appropriate strategy combining roles typical to cyberbullying, content-based, and user-based. The results showed better performance with the combined use of all features. Van Hee et al. [69] developed the corpus of Dutch social media messages and annotated the same in different categories of cyberbullying, such as threats and insults. The authors also added the comprehensive details that the participants involved in bullying (victim, cyber predator, and bystander identification). Zhao et al. [70] extended the insult to create bullying features based on word embedding and obtained an f-measure of 0.78 with an SVM classifier. In addition, the novel features were derived from a dictionary of standard terms used by neurotics in social networks. The authors in [71] have used the Word2Vec embedding model-based neural network, which was utilized to represent textual health data with a semantic context.

Moreover, unique domain ontologies are incorporated into the Word2Vec model. These ontologies provide additional details on a neural network model that recognizes the semantic sense of uncommon words. New semantic information utilizing the Bi-LSTM model is employed to precisely distinguish unstructured and structured health data. A different work is used the decision tree C4.5

classifier based on TF-IDF weighting method to detect and classify hoax news on Twitter. N-gram is also utilized to extract features to the suggested C4.5 classifiers [72]. In [73], authors have suggested a novel model that incorporates the most relevant documents, reviews, and tweets from social media and news articles. In addition, they integrated a topic2vec with Word2Vec and created a word embedding model representing each word in a document with a semantic meaning and a low-dimensional vector. The authors also used ML to classify the data using the models as mentioned earlier. Table 1 summarizes and shows the comparison results of the related studies.

As cyberbullying is considered a classification issue (i.e., categorizing an instance as offensive or non-offensive), several supervised learning algorithms have been employed in this study for the further evolution of their classification accuracy and performance in detecting cyberbullying in SM, in particular on Twitter. The classifiers adopted in the current study are as follows:

2.2.1. Logistic Regression

Logistic regression is one of the well-known techniques introduced from the field of statistics by machine learning [74]. Logistic regression is an algorithm that constructs a separate hyper-plane between two datasets utilizing the logistic function [75]. The logistic regression algorithm takes features (inputs) and produces a forecast according to the probability of a class suitable for the input. For instance, if the likelihood is ≥ 0.5 , the instance classification will be a positive class; otherwise, the prediction will be for the other class (negative class) [76], as given in Equation (1). In [77–81], logistic regression was used in the implementation of predictive cyberbullying models.

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}, \quad (1)$$

if $h_{\theta}(x) \geq 0.5$, $y = 1$ (Positive class)

and if $h_{\theta}(x) \leq 0.5$, $y = 0$ (Negative class)

As stated in [82], LR works well for the binary classification problem and functions better as data size increases. LR iteratively updates the set of parameters and attempts to minimize the error function [82].

2.2.2. Logistic Light Gradient Boosting Machine

LightGBM is one of the powerful boosting algorithms in machine learning, and it is known as a gradient boosting framework that uses a tree-based learning algorithm [83]. However, it performs better compared to XGBoost and CatBoost [84]. Gradient-based One-side Sampling (GOSS) is used in LightGBM to classify the observations used to compute the separation. The LightGBM has the primary advantage of modifying the training algorithm, which significantly increases the process [85], and leads in many cases to a more efficient model [85,86]. LightGBM has been used in many classification fields, such as online behavior detection [87] and anomalies detection in big accounting data [88].

However, LightGBM was not commonly used in the area of cyberbullying detection. Thus, in this study, we attempt to explore LightGBM in cyberbullying detection to evaluate its classification accuracy.

2.2.3. Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) is an optimization algorithm used to find parameter values (coefficients) of a function (f), which minimizes cost (cost) function [89]. SGD, in contrast, performs a parameter update for each training example $x^{(i)}$ and label $y^{(i)}$, as given in Equation (2).

$$\theta = \theta - \eta \cdot \nabla_{\theta} J(\theta; x^{(i)}; y^{(i)}), \quad (2)$$

Therefore, SGD was used in building cyberbullying prediction models in social networking platforms in [90–92]. The authors in [82] claim that SGD performs faster than NB and LR, but the error is not minimum as in LR.

2.2.4. Random Forest

Random Forest (RF) classifier is an ensemble algorithm [93] that matches multiple decision-tree classifiers on different data sub-samples, using average data to enhance predictive accuracy and control of fitting [94]. Ensemble algorithms combine more than one algorithm of the same or different kinds for classifying data [95–99]. RF was commonly used in the literature for the development of cyberbullying prediction models; examples are the studies conducted by [97–99]. Consequently, RF consists of several trees used randomly to pick the variables for the classifier data. In the following four simplified steps, the construction of the RF takes place. In the training data, N is the number of examples (cases) and M the number of attributes in the classifier.

- In the training data, N is the number of examples (cases), and M is the number of attributes in the classifier.
- Selecting random attributes produces a set of arbitrary decision tresses. For each tree, a training set is selected by selecting n times out of all existing N instances. The remaining instances in the training set are used by predicting their classes to estimate the tree's error.
- M random variables are chosen for the nodes of each tree to base the decision at that node. In the training package, the most exceptional split is determined using specific m attributes. Each tree is built entirely and not pruned, as can be done in the development of a regular tree classifier.
- This architecture produces a large number of trees. For the most common class, those decision trees vote. Such processes are denominated RFs. RF builds a model consisting of a group of tree-structured classifiers, where each tree votes for the most popular class [93]. The one selected as the output is the most highly voted class.

2.2.5. AdaBoost

Adaptive boosting (AdaBoost) is an ensemble learning method, and it is a prevalent boosting technique that was initially developed to make binary classifiers more efficacious [100,101]. It uses an iterative approach to learn from weak classifiers' errors, and transform them into strong ones. Therefore, each training observation is initially assigned equal weights. It uses several weak models and attributes higher weights to experimental misclassification observations. As the results of the definitive boundaries obtained during several iterations are combined using several low models, the accuracy of the erroneously classified observations is improved. Thus, the accuracy of the overall iteration is enhanced [102]. An example of AdaBoost classifier implantation is shown in Figure 1, where it showed a similar dataset that has two features and two classes in which weak learner #2 improve by mistake made by weak learner #1 and the accuracy of the misclassified observations is further improved when the two-week classifier are combined (strong learner).

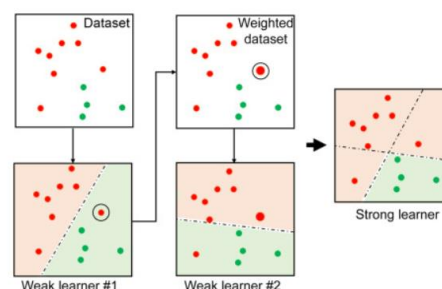


Figure 1. Implementation of Adaboost classifier [101].

Moreover, AdaBoost has been used in cyberbullying detection by some researchers like [103] and [63], as well as, the work in [104] who used it for cyberbullying detection, where they obtained an accuracy of 76.39% with AdaBoost, utilizing unigrams, comments, profile, and media information as features.

2.2.6. Multinomial Naive Bayes

Multinomial Naive Bayes (Multinomial NB) is widely used for document/text classification problems. However, in the cyberbullying detection field, NB was the most commonly used to implement cyberbullying prediction models, such as in [78] and [64,105,106].

NB classifiers were developed by applying the theorem of Bayes among features. This model assumes that a parametric model produces the text and makes use of training data to determine Bayes-optimal parameter estimates of the model. With those approximations, it categorizes produced test data [107]. NB classifiers can accommodate an arbitrary number of separate continuous or categorical functions. Assuming the functions are distinct, a task for estimating high-dimensional density is reduced to estimating one-dimensional kernel density. The NB algorithm is a learning algorithm based on the Bayes theorem's use with strong (naive) assumptions of independence. Therefore, in [108], NB was discussed in detail.

2.2.7. Support Vector Machine Classifier

Support Vector Machine (SVM) is a supervised machine learning classifier widely utilized in text classification [61]. SVM turns the original feature space into a user-defined kernel-based higher-dimensional space and then seeks support vectors for optimizing the distance (margin) between two categories. SVM originally approximates a hyperplane separating the two categories. SVM accordingly selects samples from both categories, which are nearest to the hyperplane, referred to as support vectors [109].

SVM seeks to efficiently distinguish the two categories (e.g., positive and negative). If the dataset is separable by nonlinear boundaries, specific kernels are implemented in the SVM to turn the function space appropriately. Soft margin is utilized to prevent overfitting by giving less weighting to classification errors along the decision boundaries for a dataset that is not easily separable [101]. In this research, we utilize SVM with a linear kernel for the basis function. Figure 2 shows the SVM classifier implementation for a dataset with two features and two categories where all samples for the training are depicted as circles or stars. Support vectors (referred to as stars) are for each of the two categories from the training samples, meaning that they are nearest to the hyperplane among the other training samples. Two results of the training were misclassified because they were on the wrong side of the hyperplane.

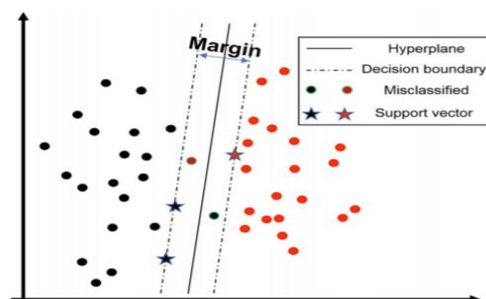


Figure 2. Implementation of Support Vector Machine (SVM) classifier [102].

Therefore, SVM was used to construct cyberbullying prediction models in [104] and found to be effective and efficient. However, the work in [61] reported that the accuracy decreased when the data size increased, suggesting that SVM may not be ideal for dealing with frequent language ambiguities typical of cyberbullying.

Table 1. Comparison of The Related Studies.

Authors	Year	Feature	Classifier	Accuracy	Dataset used
Raisi and Huang [57]	2016	N/A	N/A	They did not evaluate their proposed model	Twitter and Ask.fm datasets
Reynolds, Kontostathis and Edwards [48]	2011	Bag of Words (BoW)	Sequential Minimal Optimization (SMO), IBK, JRip, J48	The model was capable of recognizing 78.5% posts in Formspring dataset	Formspring Link: www.Formspring.me
Nahar et al. [66]	2014	TF-IDF unigrams	Ensemble	NA	MySpace, Slashdot, Kongregate, Twitter
Yin et al. [47]	2009	TF-IDF	SVM	Kongregate (0.289) Slashdot (0.273) MySpace (0.351) Correctly identify 85.3% as cyberbullying posts and 51.91% as innocent posts of MySpace dataset	MySpace, Slashdot, Kongregate
Bayzick et al. [110]	2011	Second person pronouns, Swear word,	NA		MySpace
Rafiq et al. [26]	2018	Negative comments, Total negative words, Unigrams	AdaBoost, LR	NA	Datasets of Vine
Galán-García et al. [64]	2016	TF-IDF, N-gram	NB, KNN, RF, J48, SMO	- SMO (68.47%) - J48 (65.81%) - RF (66.48%) - NB (33.91%) - KNN (59.79%)	Twitter
Al-garadi et al. [60]	2015	Unigram 3-g	SVM, NB	- Naïve Bayes (71%) - SVM (78%)	Twitter
Salminen et al. [67]	2020	TF-IDF	LR, NB, SVM, XGBoost	-LR (76.8%) - NB (60.6%) -SVM (64.8%) - XGBoost (77.4%)	A total of 197,566 comments from four platforms: YouTube, Reddit, Wikipedia, and Twitter,
Dinakar et al. [44]	2012	Profanity, BoW, TF-IDF, Weighted unigrams	J48, SVM, NB-based learner, Rule-based Jrip	-NB (63%) - J48 (61%) -SVM (72%) - Rule-based Jrip (70.39%)	Formspring, Youtube
Dadvar et al. [68]	2013	Emoticons, Message length, N-gram, Bully keywords, Pronouns	SVM	NA	YouTube
Van Hee et al. [69]	2018	Character n-gram BoW, Word n-gram BoW	LSVM	F1 score of 64% and 61% for English and Dutch respectively	Posts of ASKfm in Dutch and English
Cheng et al. [111]	2019	NA	LR, LSVM, RF	NA	Vine, Instagram
Authors in this study	2020	TF-IDF and Word2Vec	LR, LGBM, SGD, RF, AdaBoost, NB, and SVM	- LR (90.57%) - LGBM (90.55%) - SGD (90.6%) - RF (89.84%) - AdaBoost (89.30%) - NB (81.39%) - SVM (67.13%)	Twitter

3. Materials and Methods

This section describes the dataset used for cyberbullying detection on Twitter, its visualization and the proposed methodology for conducting sentiment analysis on the dataset selected, as well as discussing the evaluation metrics of each classifier used.

3.1. Dataset

Detecting cyberbullying in social media through cyberbullying keywords and using machine learning for detection are theoretical and practical challenges. From a practical perspective, the researchers are still attempting to detect and classify the offensive contents based on the learning model. However, the classification accuracy and the implementation of the right model remain a critical challenge to construct an effective and efficient cyberbullying detection model. In this study, we

used a global dataset of 37,373 tweets to evaluate seven classifiers that are commonly used in cyberbully content detection. Therefore, our dataset is taken from two sources [8,45]; and has been divided into two parts. The first part contains 70% of the tweets used for training purposes, and the other part contains 30% used for predication purpose. The evolution of each classifier will be conducted based on the performance metrics, as discussed in Section 4.

3.2. Model Overview

Figure 3 illustrates the proposed model of cyberbullying detection, where it has four phases: the preprocessing phase, the feature extraction phase, classification phase, and evaluation phase. Each phase has been discussed in detail in this section.

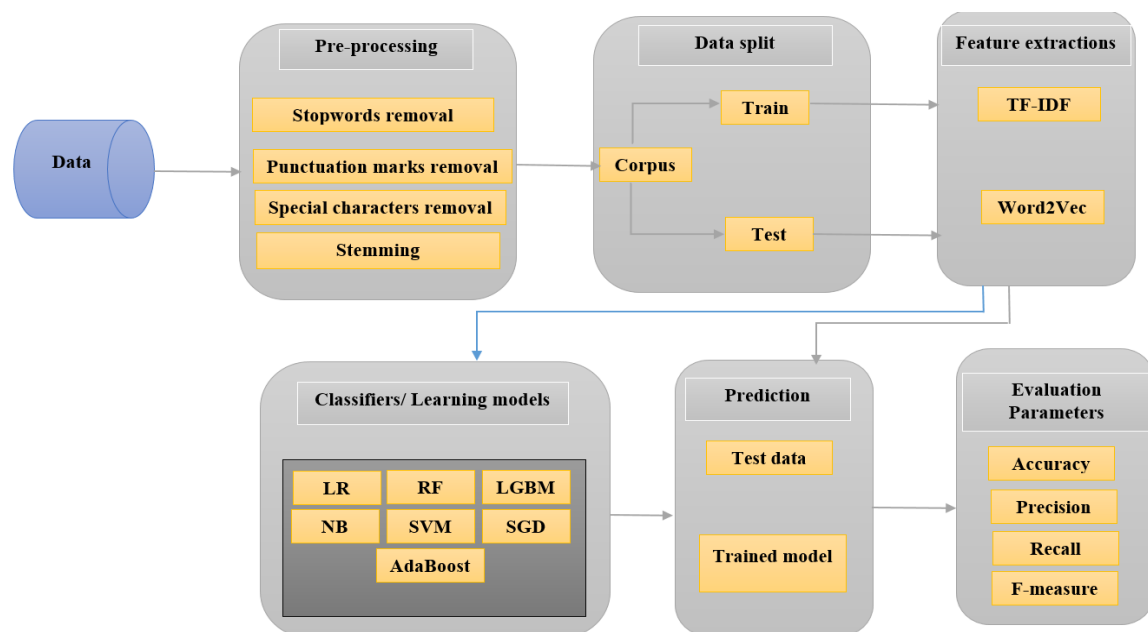


Figure 3. Overview of Cyberbullying Classification Pipeline.

3.2.1. Pre-processing

The preprocessing step is essential in cyberbullying detection. It consists of both cleaning of texts (e.g., removal of stop words and punctuation marks), as well as spam content removal [112]. In the proposed model, it has been applied to remove and clean unwanted noise in text detection. For example, stop words, special characters, and repeated words were removed. Then, the stemming for the remaining words to their original roots has been applied as a result of this preprocessing, and the dataset containing clean tweets is produced for the proposed model to be run and predicted.

3.2.2. Feature Extraction

Feature extraction is a critical step for text classification in cyberbullying. In the proposed model, we have used TF-IDF and Word2Vec techniques for feature extraction. TF-IDF is a combination of TF and IDF (term frequency-inverse document frequency), and this algorithm is based on word statistics for text feature extraction. This model considers only the expressions of words that are the same in all texts [72]. Therefore, TF-IDF is one of the most commonly used feature extraction techniques in text detection [16]. Word2Vec is a two-layer neural net that “vectorizes” words to process text. Its input is a corpus of text, and its output is a set of vectors: attribute vectors representing words in that structure [49]. The Word2Vec method uses two hidden layers of shallow neural networks, continuous bag-of-words (CBOW), and the Skip-gram model to construct a high-dimensional vector

for each word [15]. The Skip-gram model is based on a corpus of terms w and meaning c . The aim is to increase the likelihood of:

$$\operatorname{argmax}_{\theta} \prod_{w \in T} \left[\prod_{c \in C} p(c | w; \theta) \right], \quad (3)$$

where T refers to text, and θ is a parameter of $p(c | w; \theta)$. Figure 4 illustrates the Word2Vec model architecture, where CBOW model attempts to find a word based on previous terms, while Skip-gram attempts to find terms that could fall in the vicinity of each word.

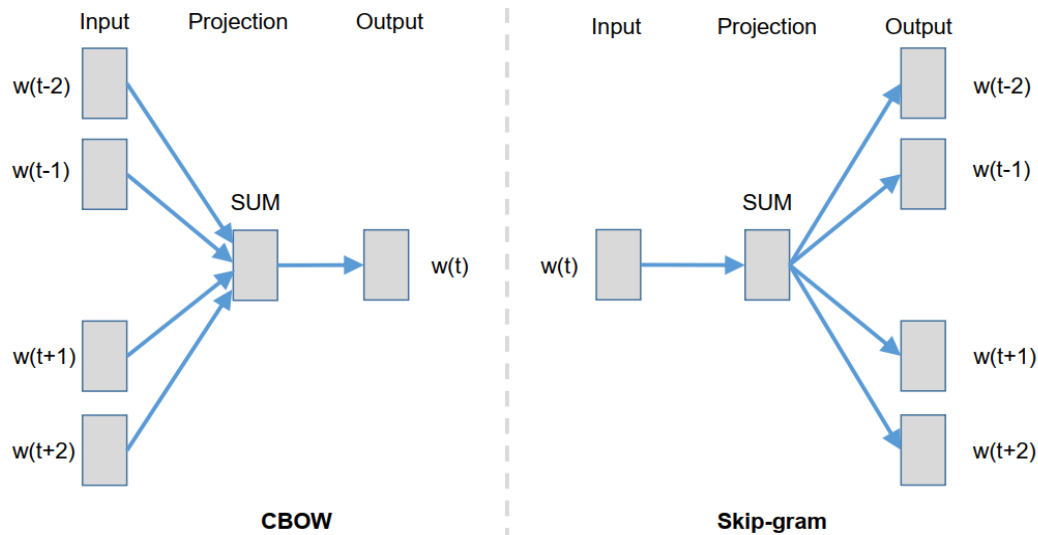


Figure 4. Continuous bag-of-words (CBOW) and Skip-gram model architecture. The Word2Vec technique implements both training models. The basic idea behind the two training models is that either a word is utilized to predict the context of it or the other way around—to use the context to predict a current word.

Utilizing TF-IDF is weighted by its relative frequency instead of merely counting the words, which would overemphasize frequent words. The TF-IDF features notify the model if a word appears more often in a statement than the entire text corpus does typically. Prior work has found TF-IDF features useful for cyberbullying detection in SM [113]. As with BOW, the TF-IDF vocabulary is constructed during model training and then reused for test prediction. Both BOW and TF-IDF are considered to be simple, proven methods for classifying text [114]. In Equation (4), the mathematical representation by TF-IDF of the weight of a term in a document is given.

$$W(d, t) = TF(d, t) * \log\left(\frac{N}{df(t)}\right), \quad (4)$$

In this case, N is the number of documents and $df(t)$ is the number of documents in the corpus containing the word t . In Equation (4), the first term enhances the recall, while the second term enhances the word embedding accuracy [52].

3.2.3. Classification Techniques

In this study, various classifiers have been used to classify whether the tweet is cyberbullying or non-cyberbullying. The classifier models constructed are LR, Light LGBM, SGD, RF, AdaBoost, naïve Bayes, and SVM. These classifiers have been discussed in Section 2, and the evaluation of their performance is carried out in Section 4.

4. Results and Discussion

This section presents the results of the experiments and discusses their significance. First, each classifier's performance results have been listed and discussed in Table 2, where it shows the evaluations of each classifier in terms of precision, recall, and F1 score, respectively. Secondly, the training time complexity of each algorithm is illustrated in Table 3. These will be discussed in detail in the following sections.

Table 2. Performance Summary of Algorithms.

No.	Algorithm	Accuracy	Precision	Recall	F1 Score	Prediction Time
1	Logistic Regression	90.57%	0.9518	0.9053	0.9280	0.0015
2	LGBM Classifier	90.55%	0.9614	0.8951	0.9271	0.0515
3	SGD Classifier	90.6%	0.9683	0.8890	0.9270	0.0016
4	Random Forest	89.84%	0.9338	0.9134	0.9235	2.5287
5	AdaBoost Classifier	89.30%	0.9616	0.8756	0.9166	0.1497
6	Multinomial NB	81.39%	0.7952	0.9736	0.8754	0.0034
7	SVM	67.13%	0.6713	1.0000	0.8033	39.9592

Table 3. Time Complexity of Algorithms.

No.	Parameters	Algorithm	Training/Prediction Time (s)
1	Best Training Time	Multinomial NB	0.014
2	Worst Prediction Time	RF	2.5287
3	Best Prediction Time	LR	0.0015
4	Worst Prediction Time	SVM	39.96

4.1. Evaluation Metrics

The effectiveness of a proposed model was examined in this study by utilizing several evaluation measures to evaluate how successfully the model can differentiate cyberbullying from non-cyberbullying. In this study, seven machine learning algorithms have been constructed, namely, LR, Light LGBM, SGD, RF, AdaBoost, Naive Bayes, and SVM. It is essential to review standard assessment metrics in the research community to understand the performance of conflicting models. The most widely used criteria for evaluating SM platforms (e.g., Twitter) with cyberbullying classifiers are as follows:

Accuracy

Accuracy calculates the ratio of the actual detected cases to the overall cases, and it has been utilized to evaluate models of cyberbullying predictions in [60,65,79]. Therefore, it can be calculated as follows:

$$\text{Accuracy} = \frac{(tp + tn)}{(tp + fp + tn + fn)} \quad (5)$$

where *tp* means true positive, *tn* is a true negative, *fp* denotes false positive, and *fn* is a false negative.

- Precision calculates the proportion of relevant tweets among true positive (*tp*) and false positive (*fp*) tweets belonging to a specific group.
- Recall calculates the ratio of retrieved relevant tweets over the total number of relevant tweets.
- F-Measure provides a way to combine precision and recall into a single measure that captures both properties.

The three evaluation measures listed above have been utilized to evaluate cyberbullying prediction models in [67,79,98,104]. They are calculated as follows:

$$\text{Precision} = \frac{tp}{(tp + fp)}, \quad (6)$$

$$\text{Recall} = \frac{tp}{(tp + fn)}, \quad (7)$$

$$\text{F measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \quad (8)$$

4.2. Performance Result of Classifiers

The proposed model utilizes the selected seven ML classifiers with two different feature extraction techniques. These techniques were set empirically to achieve higher accuracy. For instance, LR achieved the best accuracy and F1 score in our dataset, where the classification accuracy and F1 score are 90.57% and 0.9280, respectively. Meanwhile, there is a slight difference between LR, SGD, and LGBM classifier performance, where SGD achieved an accuracy of 90.6%, but the F1 score was lower than LR. However, the LGBM classifier achieved an accuracy of 90.55%, and the F1 score was 0.9271. This means LR performs better than other classifiers, as shown in Table 2.

Moreover, RF and AdaBoost have achieved almost the same accuracy, but in terms of F1 Score, RF performs better than AdaBoost. Multinomial NB has achieved low accuracy and precision with a detection rate of 81.39% and 0.7952, respectively, and we can notice that the excellent recall levels-out the low precision, giving a good F-measure score of 0.8754 as illustrated in Table 2.

Finally, SVM has achieved the lowest accuracy and precision in our dataset, as shown in Figure 5. Nevertheless, it achieved the best recall compared to the rest of the classifiers implemented in the current research. Furthermore, some studies have looked at the automatic cyberbullying detection incidents; for example, an effect analysis based on lexicon and SVM was found to be effective in detecting cyberbullying. However, the accuracy decreased when data size increased, suggesting that SVM may not be ideal for dealing with common language ambiguities typical of cyberbullying [61]. This proves that the low accuracy achieved by SVM is due to the large dataset used in this research.

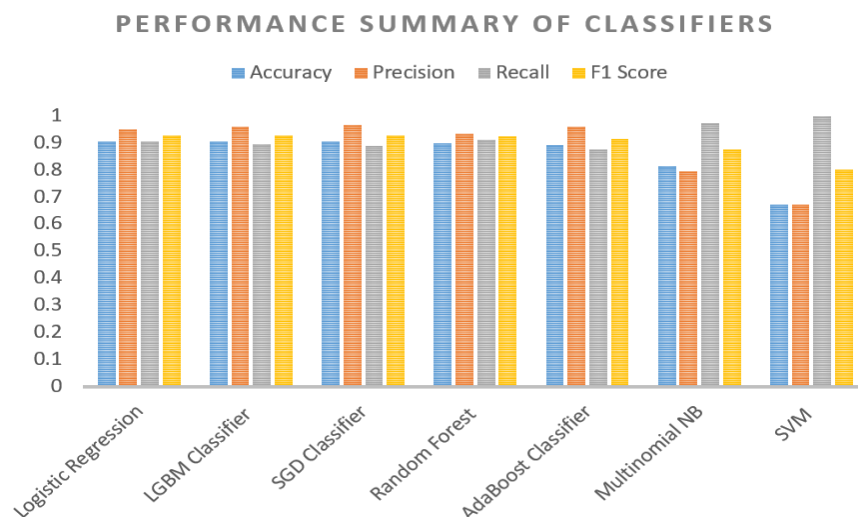


Figure 5. Performance Summary of Algorithms.

F-measure is one of the most effective evaluation metrics. In this research, the seven classifiers' performances were computed using the F-measure metric, as shown in Figure 6. Furthermore, the performances of all ML classifiers are enhanced by producing additional data utilizing data

synthesizing techniques. Multinomial NB assumes that every function is independent, but this is not true in real situations [115]. Therefore, it does not outperform LR in our research as well. As stated in [116], LR performs well for the binary classification problem and works better as data size increases. LR updates several parameters iteratively and tries to eliminate the error. Simultaneously, SGD uses a single sample and uses a similar approximation to update the parameters. Therefore, SGD performs almost as LR, but the error is not as reduced as in LR [92]. Consequently, it is not surprising that LR also outperforms the other classifiers in our study.

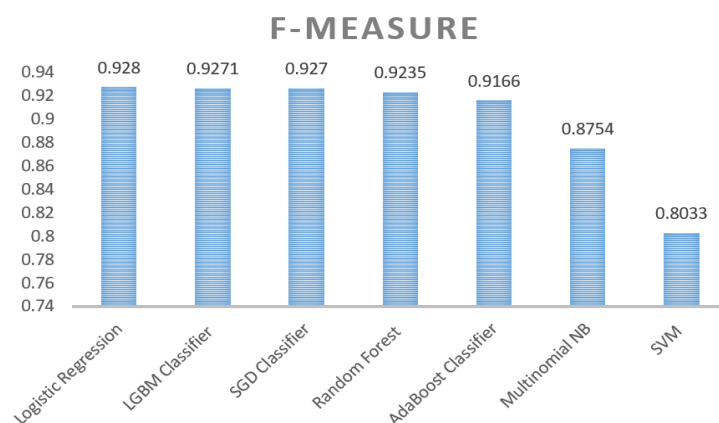


Figure 6. Comparison between the Best Classifiers in Terms of F-Measure.

4.3. Time Complexity of Algorithms

Table 3 shows the time complexity of the best and the worst algorithms in terms of training and prediction time. The results in Table 3 indicate that Multinomial NB has achieved the best training time, and RF has obtained the worst training time, 0.014s and 2.5287s, respectively. Meanwhile, LR outperforms all the classifiers implemented in this research. However, there were slight differences between SGD and Multinomial NB compared to LR, as shown in Table 3.

5. Conclusions

Cyberbullying has become a severe problem in modern societies. This paper proposed a cyber-bully detection model whereby several classifiers based on TF-IDF and Word2Vec feature extraction have been used. Furthermore, various methods of text classification based on machine learning were investigated. The experiments were conducted on a global Twitter dataset. The experimental results indicate that LR achieved the best accuracy and F1 score in our dataset, where the classification accuracy and F1 score are 90.57% and 0.9280, respectively.

Meanwhile, there is a slight difference between LR, SGD, and LGBM classifier performance, where SGD achieved an accuracy of 90.6%, but the F1 score was lower than LR. However, the LGBM classifier achieved an accuracy of 90.55%, and the F1 score was 0.9271. This means that LR performs better than other classifiers. Moreover, during the experiments, it was observed that LR performs better as data size increases and obtains the best prediction time compared to other classifiers used in this study. Therefore, SGD performs almost as LR, but the error is not minimal as in LR.

The feature extraction is a critical aspect in machine learning to enhance the detection accuracy. In this paper, we did not investigate many feature extraction techniques. Thus, one of the improvements is to incorporate and test different feature extractions to improve the detection rate of both classifiers LR and SGD. Another limitation that we are working on is building a real-time cyberbully detection platform, which will be useful to instantly detect and prevent the cyberbully. Another research direction is working on cyberbully detection in various languages, mainly in an Arabic context.

Author Contributions: Conceptualization, methodology, validation, formal analysis, investigation and visualization, A.M. and S.M.F.; software and writing—original draft preparation, A.M.; writing—review and editing, S.M.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Prince Sultan University, Riyadh, Saudi Arabia.

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

References

- Edosomwan, S.; Prakasan, S.K.; Kouame, D.; Watson, J.; Seymour, T. The history of social media and its impact on business. *J. Appl. Manag. Entrep.* **2011**, *16*, 79–91.
- Bauman, S. *Cyberbullying: What Counselors Need to Know*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
- Pereira-Kohatsu, J.C.; Quijano-Sánchez, L.; Liberatore, F.; Camacho-Collados, M. Detecting and Monitoring Hate Speech in Twitter. *Sensors* **2019**, *19*, 4654. [\[CrossRef\]](#)
- Miller, K. Cyberbullying and its consequences: How cyberbullying is contorting the minds of victims and bullies alike, and the law's limited available redress. *S. Cal. Interdisc. Law J.* **2016**, *26*, 379.
- Price, M.; Dalgleish, J. Cyberbullying: Experiences, impacts and coping strategies as described by australian young people. *Youth Stud. Aust.* **2010**, *29*, 51.
- Smith, P.K. Cyberbullying and Cyber Aggression. In *Handbook of School Violence and School Safety*; Informa UK Limited: Colchester, UK, 2015.
- Sampasa-Kanyinga, H.; Roumeliotis, P.; Xu, H. Associations between Cyberbullying and School Bullying Victimization and Suicidal Ideation, Plans and Attempts among Canadian Schoolchildren. *PLoS ONE* **2014**, *9*, e102145. [\[CrossRef\]](#)
- Davidson, T.; Warmesley, D.; Macy, M.; Weber, I. Automated Hate Speech Detection and the Problem of Offensive Language. *arXiv* **2017**, arXiv:1703.04009.
- Mc Guckin, C.; Corcoran, L. (Eds.) *Cyberbullying: Where Are We Now? A Cross-National Understanding*; MDPI: Wuhan, China, 2017.
- Vaillancourt, T.; Faris, R.; Mishna, F. Cyberbullying in Children and Youth: Implications for Health and Clinical Practice. *Can. J. Psychiatry* **2016**, *62*, 368–373. [\[CrossRef\]](#)
- Görzig, A.; Ólafsson, K. What Makes a Bully a Cyberbully? Unravelling the Characteristics of Cyberbullies across Twenty-Five European Countries. *J. Child. Media* **2013**, *7*, 9–27. [\[CrossRef\]](#)
- Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **1988**, *24*, 513–523.
- Liu, Q.; Wang, J.; Zhang, D.; Yang, Y.; Wang, N. Text Features Extraction based on TF-IDF Associating Semantic. In Proceedings of the 2018 IEEE 4th International Conference on Computer and Communications (ICCC), Chengdu, China, 7–10 December 2018; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2018; pp. 2338–2343.
- Pennington, J.; Socher, R.; Manning, C. Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014.
- Goldberg, Y.; Levy, O. word2vec Explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv* **2014**, arXiv:1402.3722.
- Li, J.; Huang, G.; Fan, C.; Sun, Z.; Zhu, H. Key word extraction for short text via word2vec, doc2vec, and textrank. *Turk. J. Electr. Eng. Comput. Sci.* **2019**, *27*, 1794–1805. [\[CrossRef\]](#)
- Jiang, C.; Zhang, H.; Ren, Y.; Han, Z.; Chen, K.-C.; Hanzo, L. Machine Learning Paradigms for Next-Generation Wireless Networks. *IEEE Wirel. Commun.* **2016**, *24*, 98–105. [\[CrossRef\]](#)
- Medhat, W.; Hassan, A.; Korashy, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Eng. J.* **2014**, *5*, 1093–1113. [\[CrossRef\]](#)
- Al-Garadi, M.A.; Hussain, M.R.; Khan, N.; Murtaza, G.; Nweke, H.F.; Ali, I.; Mujtaba, G.; Chiroma, H.; Khattak, H.A.; Gani, A. Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges. *IEEE Access* **2019**, *7*, 70701–70718. [\[CrossRef\]](#)
- Maalouf, M. Logistic regression in data analysis: An overview. *Int. J. Data Anal. Tech. Strat.* **2011**, *3*, 281–299. [\[CrossRef\]](#)

21. Hosmer, D.W.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*; Wiley: Hoboken, NJ, USA, 2013; Volume 398.
22. Chavan, V.S.; Shylaja, S.S. Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In Proceedings of the 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Kochi, India, 10–13 August 2015; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2015; pp. 2354–2358.
23. Mangaonkar, A.; Hayrapetian, A.; Raje, R. Collaborative detection of cyberbullying behavior in Twitter data. In Proceedings of the 2015 IEEE International Conference on Electro/Information Technology (EIT), Dekalb, IL, USA, 21–23 May 2015; IEEE: New York, NY, USA, 2015; pp. 611–616.
24. Leon-Paredes, G.A.; Palomeque-Leon, W.F.; Gallegos-Segovia, P.L.; Vintimilla-Tapia, P.E.; Bravo-Torres, J.F.; Barbosa-Santillan, L.I.; Paredes-Pinos, M.M. Presumptive Detection of Cyberbullying on Twitter through Natural Language Processing and Machine Learning in the Spanish Language. In Proceedings of the 2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), Valparaiso, Chile, 13–27 November 2019; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2019; pp. 1–7.
25. Ho, S.M.; Li, W.; Lai, C.J.; Ankamah, B. Charged Language on Twitter: A Predictive Model of Cyberbullying to Prevent Victimization. In Proceedings of the 2019 AIS SIGSEC Special Interest Group 10th Annual Workshop on Information Security and Privacy (WISP), International Conference on Information Systems (ICIS), AIS, Munich, Germany, 15 December 2019; pp. 1–10.
26. Ibn Rafiq, R.; Hosseinmardi, H.; Han, R.; Lv, Q.; Mishra, S. Scalable and timely detection of cyberbullying in online social networks. In Proceedings of the 33rd Annual ACM Symposium on Applied Computing—SAC '18, Pau, France, 9–13 April 2018; pp. 1738–1747.
27. Cheng, L.; Li, J.; Silva, Y.N.; Hall, D.L.; Liu, H. XBully: Cyberbullying Detection within a Multi-Modal Context. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, Melbourne, Australia, 11–15 February 2019; ACM: New York, NY, USA, 2019; pp. 339–347.
28. Nahar, V.; Li, X.; Zhang, H.L.; Pang, C. Detecting cyberbullying in social networks using multi-agent system. *Web Intell. Agent Syst. Int. J.* **2014**, *12*, 375–388. [CrossRef]
29. Mandot, P. What Is Lightgbm, How to Implement It? How to Fine Tune the Parameters? Medium. 2017. Available online: <https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc> (accessed on 22 July 2020).
30. Rahman, S.; Irfan, M.; Raza, M.; Ghori, K.M.; Yaqoob, S.; Awais, M. Performance Analysis of Boosting Classifiers in Recognizing Activities of Daily Living. *Int. J. Environ. Res. Public Health* **2020**, *17*, 1082. [CrossRef]
31. Brownlee, J. Gradient Boosting With Scikit-Learn, Xgboost, Lightgbm, and Catboost. Machine Learning Mastery. 2020. Available online: <https://machinelearningmastery.com/gradient-boosting-with-scikit-learn-xgboost-lightgbm-and-catboost/> (accessed on 22 July 2020).
32. Zinovyeva, E.; Härdle, W.K.; Lessmann, S. Antisocial online behavior detection using deep learning. *Decis. Support Syst.* **2020**, *138*, 113362. [CrossRef]
33. Bhattacharya, I.; Lindgreen, E.R. A Semi-Supervised Machine Learning Approach to Detect Anomalies in Big Accounting Data. In Proceedings of the ECIS, Marrakech, Morocco, 15–17 June 2020.
34. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*; Morgan Kaufmann Publishers Inc.: Burlington, MA, USA, 2017; pp. 3146–3154.
35. Brownlee, J. *Master Machine Learning Algorithms: Discover How They Work and Implement Them from Scratch*; Machine Learning Mastery: Vermont, Australia, 2016.
36. Pawar, R.; Agrawal, Y.; Joshi, A.; Gorrepati, R.; Raje, R.R. Cyberbullying detection system with multiple server configurations. In Proceedings of the 2018 IEEE International Conference on Electro/Information Technology (EIT), Rochester, MI, USA, 3–5 May 2018; IEEE: New York, NY, USA, 2018; pp. 0090–0095.
37. Aci, C.; Çürük, E.; Eşsiz, E.S. Automatic Detection of Cyberbullying in FORMSPRING.Me, Myspace and Youtube Social Networks. *Turk. J. Eng.* **2019**, *3*, 168–178. [CrossRef]
38. Pawar, R.; Raje, R.R. Multilingual Cyberbullying Detection System. In Proceedings of the 2019 IEEE International Conference on Electro Information Technology (EIT), Brookings, SD, USA, 20–22 May 2019; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2019; pp. 040–044.

39. Why Logistic Regression over Naïve Bayes. Available online: https://medium.com/@sangha_deb/naive-bayes-vs-logisticregression-a319b07a5d4c (accessed on 22 July 2020).
40. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
41. Patel, S. Chapter 5: Random Forest Classifier. 2017. Available online: <https://medium.com/machine-learning-101/chapter-5-random-forest-classifier-56dc7425c3e1> (accessed on 22 July 2020).
42. Louppe, G. Understanding random forests: From theory to practice. *arXiv* **2014**, arXiv:1407.7502.
43. Novalita, N.; Herdiani, A.; Lukmana, I.; Puspandari, D. Cyberbullying identification on twitter using random forest classifier. *J. Physics Conf. Ser.* **2019**, *1192*, 012029. [[CrossRef](#)]
44. García-Recuero, Á. Discouraging Abusive Behavior in Privacy-Preserving Online Social Networking Applications. In Proceedings of the 25th International Conference Companion on World Wide Web—WWW '16 Companion, Montreal, QC, Canada, 11–15 April 2016; Association for Computing Machinery (ACM): New York, NY, USA, 2016; pp. 305–309.
45. Waseem, Z.; Hovy, D. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In Proceedings of the NAACL Student Research Workshop, San Diego, CA, USA, 16–17 June 2016; Association for Computational Linguistics (ACL): Stroudsburg, PA, USA, 2016; pp. 88–93.
46. Chengsheng, T.; Huacheng, L.; Bing, X. AdaBoost typical Algorithm and its application research. In *MATEC Web of Conferences*; EDP Sciences: Ulis, France, 2017; Volume 139, p. 00222.
47. Chatterjee, R.; Datta, A.; Sanyal, D.K. Ensemble Learning Approach to Motor Imagery EEG Signal Classification. In *Machine Learning in Bio-Signal Analysis and Diagnostic Imaging*; Elsevier BV: Amsterdam, The Netherlands, 2019; pp. 183–208.
48. Misra, S.; Li, H. Noninvasive fracture characterization based on the classification of sonic wave travel times. In *Machine Learning for Subsurface Characterization*; Elsevier BV: Amsterdam, The Netherlands, 2020; pp. 243–287.
49. Ibn Rafiq, R.; Hosseinmardi, H.; Han, R.; Lv, Q.; Mishra, S.; Mattson, S.A. *Careful What You Share in Six Seconds*. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015—ASONAM '15*; Association for Computing Machinery (ACM): New York, NY, USA, 2015; pp. 617–622.
50. Tarwani, S.; Jethanandani, M.; Kant, V. Cyberbullying Detection in Hindi-English Code-Mixed Language Using Sentiment Classification. In *Communications in Computer and Information Science*; Springer Science and Business Media LLC: Singapore, 2019; pp. 543–551.
51. Raza, M.O.; Memon, M.; Bhatti, S.; Bux, R. Detecting Cyberbullying in Social Commentary Using Supervised Machine Learning. In *Advances in Intelligent Systems and Computing*; Springer Science and Business Media LLC: Singapore, 2020; pp. 621–630.
52. Galán-García, P.; De La Puerta, J.G.; Gómez, C.L.; Santos, I.; Bringas, P.G. Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. *Log. J. IGPL* **2015**, *24*, jzv048. [[CrossRef](#)]
53. Akhter, A.; Uzzal, K.A.; Polash, M.A. Cyber Bullying Detection and Classification using Multinomial Naïve Bayes and Fuzzy Logic. *Int. J. Math. Sci. Comput.* **2019**, *5*, 1–12. [[CrossRef](#)]
54. Nandakumar, V. Cyberbullying revelation in twitter data using naïve bayes classifier algorithm. *Int. J. Adv. Res. Comput. Sci.* **2018**, *9*, 510–513. [[CrossRef](#)]
55. Dinakar, K.; Reichart, R.; Lieberman, H. Modeling the detection of textual cyberbullying. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 17–21 July 2011.
56. Chen, Y.; Zhou, Y.; Zhu, S.; Xu, H. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing, Amsterdam, The Netherlands, 3–5 September 2012; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2012; pp. 71–80.
57. Sintaha, M.; Satter, S.B.; Zawad, N.; Swarnaker, C.; Hassan, A. Cyberbullying Detection Using Sentiment Analysis in Social Media. Ph.D. Thesis, BRAC University, Dhaka, Bangladesh, 2016.
58. McCallum, A.; Nigam, K. A comparison of event models for naïve bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*; Amer Assn for Artificial Intelligence: Cambridge, MA, USA, 1998; Volume 752, pp. 41–48.
59. Buczak, A.L.; Guven, E. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Commun. Surv. Tutor.* **2015**, *18*, 1153–1176. [[CrossRef](#)]

60. Zhang, H. Exploring conditions for the optimality of naïve bayes. *Int. J. Pattern Recognit. Artif. Intell.* **2005**, *19*, 183–198. [\[CrossRef\]](#)
61. Joachims, T. Text categorization with Support Vector Machines: Learning with many relevant features. In *the Computer Vision—ECCV 2018*; Springer Science and Business Media LLC: Berlin, Germany, 1998; pp. 137–142.
62. Ptaszynski, M.; Eronen, J.K.K.; Masui, F. Learning Deep on Cyberbullying is Always Better than Brute Force. In *Proceedings of the LaCATODA@IJCAI*, Melbourne, Australia, 21 August 2017; pp. 3–10.
63. Chatzakou, D.; Leontiadis, I.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; Vakali, A.; Kourtellis, N. Detecting Cyberbullying and Cyberaggression in Social Media. *ACM Trans. Web* **2019**, *13*, 1–51. [\[CrossRef\]](#)
64. Irena, B.; Setiawan, E.B. Fake News (Hoax) Identification on Social Media Twitter using Decision Tree C4.5 Method. *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)* **2020**, *4*, 711–716. [\[CrossRef\]](#)
65. Hosseinmardi, H.; Mattson, S.A.; Ibn Rafiq, R.; Han, R.; Lv, Q.; Mishra, S. Detection of Cyberbullying Incidents on the Instagram Social Network. *arXiv* **2015**, arXiv:1503.03909.
66. Ptaszynski, M.; Dybala, P.; Matsuba, T.; Masui, F.; Rzepka, R.; Araki, K. Machine learning and affect analysis against cyber-bullying. In *Proceedings of the the 36th AISB*, Leicester, UK, 29 March–1 April 2010; pp. 7–16.
67. Dadvar, M.; Jong, F.D.; Ordelman, R.; Trieschnigg, D. Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*, University of Ghent, Gent, Belgium, 23–24 February 2012.
68. Van Hee, C.; Lefever, E.; Verhoeven, B.; Mennes, J.; Desmet, B.; De Pauw, G.; Daelemans, W.; Hoste, V. Detection and fine-grained classification of cyberbullying events. In *International Conference Recent Advances in Natural Language Processing (RANLP)*; INCOMA Ltd.: Shoumen, Bulgaria, 2015; pp. 672–680.
69. Kowsari, K.; Meimandi, K.J.; Heidarysafa, M.; Mendu, S.; Barnes, L.; Brown, D. Text Classification Algorithms: A Survey. *Information* **2019**, *10*, 150. [\[CrossRef\]](#)
70. Sahlgren, M.; Isbister, T.; Olsson, F. Learning Representations for Detecting Abusive Language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, Brussels, Belgium, 31 October–1 November 2018; pp. 115–123.
71. Salminen, J.; Almerexhi, H.; Milenković, M.; Jung, S.G.; An, J.; Kwak, H.; Jansen, B.J. Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Twelfth International AAAI Conference on Web and Social Media*; AAAI Press: Palo Alto, CA, USA, 2018.
72. Ottesen, C. Comparison between Naïve Bayes and Logistic Regression. DataEspresso. 2017. Available online: <https://dataespresso.com/en/2017/10/24/comparison-between-naive-bayes-and-logistic-regression/#:~:text=Na%C3%AFve%20Bayes%20has%20a%20naive,belonging%20to%20a%20certain%20class> (accessed on 24 July 2020).
73. Deb, S. Naive Bayes vs. Logistic Regression. Medium. 2016. Available online: https://medium.com/@sangha_deb/naive-bayes-vs-logistic-regression-a319b07a5d4c (accessed on 24 July 2020).
74. Snakenborg, J.; Van Acker, R.; Gable, R.A. Cyberbullying: Prevention and Intervention to Protect Our Children and Youth. *Prev. Sch. Fail. Altern. Educ. Child. Youth* **2011**, *55*, 88–95. [\[CrossRef\]](#)
75. Patchin, J.W.; Hinduja, S. Traditional and Nontraditional Bullying Among Youth: A Test of General Strain Theory. *Youth Soc.* **2011**, *43*, 727–751. [\[CrossRef\]](#)
76. Tenenbaum, L.S.; Varjas, K.; Meyers, J.; Parris, L. Coping strategies and perceived effectiveness in fourth through eighth grade victims of bullying. *Sch. Psychol. Int.* **2011**, *32*, 263–287. [\[CrossRef\]](#)
77. Olweus, D. Invited expert discussion paper Cyberbullying: An overrated phenomenon? *Eur. J. Dev. Psychol.* **2012**, *9*, 1–19. [\[CrossRef\]](#)
78. Hinduja, S.; Patchin, J.W. Cyberbullying: An Exploratory Analysis of Factors Related to Offending and Victimization. *Deviant Behav.* **2008**, *29*, 129–156. [\[CrossRef\]](#)
79. Hemphill, S.A.; Kotevski, A.; Tollit, M.; Smith, R.; Herrenkohl, T.I.; Toumbourou, J.W.; Catalano, R.F. Longitudinal Predictors of Cyber and Traditional Bullying Perpetration in Australian Secondary School Students. *J. Adolesc. Health* **2012**, *51*, 59–65. [\[CrossRef\]](#)
80. Casas, J.A.; Del Rey, R.; Ortega-Ruiz, R. Bullying and cyberbullying: Convergent and divergent predictor variables. *Comput. Hum. Behav.* **2013**, *29*, 580–587. [\[CrossRef\]](#)
81. Ang, R.P.; Goh, D.H. Cyberbullying among Adolescents: The Role of Affective and Cognitive Empathy, and Gender. *Child Psychiatry Hum. Dev.* **2010**, *41*, 387–397. [\[CrossRef\]](#)

82. Barlińska, J.; Szuster, A.; Winiewski, M. Cyberbullying among Adolescent Bystanders: Role of the Communication Medium, Form of Violence, and Empathy. *J. Community Appl. Soc. Psychol.* **2012**, *23*, 37–51. [CrossRef]
83. Ybarra, M.L.; Mitchell, K.J.; Wolak, J.; Finkelhor, D. Examining Characteristics and Associated Distress Related to Internet Harassment: Findings from the Second Youth Internet Safety Survey. *Pediatrics* **2006**, *118*, e1169–e1177. [CrossRef] [PubMed]
84. Smith, P.K.; Mahdavi, J.; Carvalho, M.; Fisher, S.; Russell, S.; Tippett, N. Cyberbullying: Its nature and impact in secondary school pupils. *J. Child Psychol. Psychiatry* **2008**, *49*, 376–385. [CrossRef]
85. Raisi, E.; Huang, B. Cyberbullying Identification Using Participant-Vocabulary Consistency. *arXiv* **2016**, arXiv:1606.08084.
86. Van Der Zwaan, J.M.; Dignum, V.; Jonker, C.M. A Conversation Model Enabling Intelligent Agents to Give Emotional Support. In *Uncertainty Theory*; Springer Science and Business Media LLC: Berlin, Germany, 2012; Volume 431, pp. 47–52.
87. Bosse, T.; Stam, S. A Normative Agent System to Prevent Cyberbullying. In Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Lyon, France, 22–27 August 2011; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2011; Volume 2, pp. 425–430.
88. Reynolds, K.; Kontostathis, A.; Edwards, L. Using Machine Learning to Detect Cyberbullying. In Proceedings of the 2011 10th International Conference on Machine Learning and Applications and Workshops, Honolulu, HI, USA, 18–21 December 2011; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2011; Volume 2, pp. 241–244.
89. Rybníček, M.; Poisel, R.; Tjoa, S. Facebook Watchdog: A Research Agenda for Detecting Online Grooming and Bullying Activities. In Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics, Manchester, UK, 13–16 October 2013; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2013; pp. 2854–2859.
90. Yin, D.; Xue, Z.; Hong, L.; Davison, B.D.; Kontostathis, A.; Edwards, L. Detection of harassment on web 2.0. In Proceedings of the Content Analysis in the WEB, Madrid, Spain, 21 April 2009; Volume 2, pp. 1–7.
91. Bayzick, J.; Kontostathis, A.; Edwards, L. Detecting the Presence of Cyberbullying Using Computer Software. 2011. Available online: <https://april-edwards.me/BayzickHonors.pdf> (accessed on 29 October 2020).
92. Al-garadi, M.A.; Varathan, K.D.; Ravana, S.D. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Comput. Hum. Behav.* **2016**, *63*, 433–443. [CrossRef]
93. Salminen, J.; Hopf, M.; Chowdhury, S.A.; Jung, S.-G.; Almerikhi, H.; Jansen, B.J. Developing an online hate classifier for multiple social media platforms. *Hum. Cent. Comput. Inf. Sci.* **2020**, *10*, 1–34. [CrossRef]
94. Dinakar, K.; Jones, B.; Havasi, C.; Lieberman, H.; Picard, R. Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying. *ACM Trans. Interact. Intell. Syst.* **2012**, *2*, 1–30. [CrossRef]
95. Dadvar, M.; Trieschnigg, R.B.; Ordelman, R.J.; De Jong, F.M. Improving Cyberbullying Detection with User Context. In *European Conference on Information Retrieval*; Springer Science and Business Media LLC: Berlin, Germany, 2013; pp. 693–696.
96. Van Hee, C.; Jacobs, G.; Emmery, C.; Desmet, B.; Lefever, E.; Verhoeven, B.; De Pauw, G.; Daelemans, W.; Hoste, V. Automatic detection of cyberbullying in social media text. *PLoS ONE* **2018**, *13*, e0203794. [CrossRef]
97. Zhao, R.; Zhou, A.; Mao, K. Automatic detection of cyberbullying on social networks based on bullying features. In Proceedings of the 17th International Conference on Distributed Computing and Networking—ICDCN '16, Singapore, 4–7 January 2016; Association for Computing Machinery (ACM): New York, NY, USA, 2016; p. 43.
98. Ahlfors Many Sources. One Theme: Analysis of Cyberbullying Prevention and Intervention Websites. *J. Soc. Sci.* **2010**, *6*, 515–522. [CrossRef]
99. Lenhart, A.; Purcell, K.; Smith, A.; Zickuhr, K. Social Media & Mobile Internet Use among Teens and Young Adults. Available online: <http://samaritanbehavioralhealth.net/files/social-media-young-adults.pdf> (accessed on 28 October 2020).
100. Webb, M.; Burns, J.; Collin, P. Providing online support for young people with mental health difficulties: Challenges and opportunities explored. *Early Interv. Psychiatry* **2008**, *2*, 108–113. [CrossRef] [PubMed]

101. Havas, J.; De Nooijer, J.; Crutzen, R.; Feron, F.J.M. Adolescents' views about an internet platform for adolescents with mental health problems. *Health Educ.* **2011**, *111*, 164–176. [CrossRef]
102. Jacobs, N.C.; Völlink, T.; Dehue, F.; Lechner, L. Online Pestkoppenstoppen: Systematic and theory-based development of a web-based tailored intervention for adolescent cyberbully victims to combat and prevent cyberbullying. *BMC Public Health* **2014**, *14*, 396. [CrossRef] [PubMed]
103. KiVa Program. Kiva Is an Anti-Bullying Programme|Kiva Antibullying Program|Just Another Kiva Koulu Site. 2020. Available online: <http://www.kivaprogram.net> (accessed on 17 August 2020).
104. Nonauharcèlement.Education.gouv.fr. Non Au Harcèlement—Appelez Le 3020. 2020. Available online: <https://www.nonauharcèlement.education.gouv.fr/> (accessed on 18 August 2020).
105. Veiligonline.be. Cyberpesten|Veilig Online. 2020. Available online: <https://www.veiligonline.be/cyberpesten> (accessed on 18 August 2020).
106. Stauffer, S.; Heath, M.A.; Coyne, S.M.; Ferrin, S. High school teachers' perceptions of cyberbullying prevention and intervention strategies. *Psychol. Sch.* **2012**, *49*, 352–367. [CrossRef]
107. Notar, C.E.; Padgett, S.; Roden, J. Cyberbullying: Resources for Intervention and Prevention. *Univers. J. Educ. Res.* **2013**, *1*, 133–145.
108. Fanti, K.A.; Demetriou, A.G.; Hawa, V.V. A longitudinal study of cyberbullying: Examining risk and protective factors. *Eur. J. Dev. Psychol.* **2012**, *9*, 168–181. [CrossRef]
109. Ybarra, M.L.; Mitchell, K.J. Prevalence and Frequency of Internet Harassment Instigation: Implications for Adolescent Health. *J. Adolesc. Health* **2007**, *41*, 189–195. [CrossRef]
110. Aricak, T.; Siyahhan, S.; Uzunhasanoglu, A.; Saribeyoglu, S.; Ciplak, S.; Yilmaz, N.; Memmedov, C. Cyberbullying among Turkish Adolescents. *CyberPsychology Behav.* **2008**, *11*, 253–261. [CrossRef]
111. Kontostathis, A.; Reynolds, K.; Garron, A.; Edwards, L. Detecting cyberbullying: Query terms and techniques. In Proceedings of the 5th Annual ACM Web Science Conference, New York, NY, USA, 23–26 June 2013; pp. 195–204.
112. Xu, J.M.; Jun, K.S.; Zhu, X.; Bellmore, A. 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, Montréal, QC, Canada, 3–8 June 2012; pp. 656–666.
113. Chowdhary, K.R. Natural language processing for word sense disambiguation and information extraction. *arXiv* **2020**, arXiv:2004.02256.
114. Liddy, E.D. Natural language processing. In *Encyclopedia of Library and Information Science*; Springer: New Delhi, India, 2001.
115. Ali, F.; El-Sappagh, S.; Islam, S.R.; Ali, A.; Attique, M.; Imran, M.; Kwak, K.-S. An intelligent healthcare monitoring framework using wearable sensors and social networking data. *Futur. Gener. Comput. Syst.* **2020**, *114*, 23–43. [CrossRef]
116. Ali, F.; Kwak, D.; Khan, P.; El-Sappagh, S.; Ali, A.; Ullah, S.; Kim, K.H.; Kwak, K.-S. Transportation sentiment analysis using word embedding and ontology-based topic modeling. *Knowl. Based Syst.* **2019**, *174*, 27–42. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

This is not a