Received April 21, 2019, accepted May 14, 2019, date of publication May 22, 2019, date of current version June 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2918354

Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges

MOHAMMED ALI AL-GARADI[1], MOHAMMAD RASHID HUSSAIN[2], NAWSHER KHAN[2], GHULAM MURTAZA[1,3], HENRY FRIDAY NWEKE[1], IHSAN ALI[1], GHULAM MUJTABA[1,3], HARUNA CHIROMA[4], HASAN ALI KHATTAK[5], AND ABDULLAH GANI[1]

[1]Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur 50603, Malaysia
[2]Collage of Computer Science, King Khalid University, Abha 61421, Saudi Arabia
[3]Department of Computer Science, Sukkur IBA University, Sukkur 65203, Pakistan
[4]Department of Computer Science, Federal College of Education (Technical), Gombe 234, Nigeria
[5]Department of Computer Science, COMSATS University Islamabad, Islamabad 45000, Pakistan

Corresponding authors: Mohammed Ali Al-Garadi (mohammedali@siswa.um.edu.my), Ihsan Ali (ihsanalichd@siswa.um.edu.my), and Ghulam Mujtaba (mujtaba@iba-suk.edu.pk)

This work was supported in part by the Deanship of Scientific Research, King Khalid University, through Research Group Project under Grant R.G.P. 1/166/40, and in part by the University of Malaya Postgraduate Research under Grant PG035-2016A.

ABSTRACT Prior to the innovation of information communication technologies (ICT), social interactions evolved within small cultural boundaries such as geo spatial locations. The recent developments of communication technologies have considerably transcended the temporal and spatial limitations of traditional communications. These social technologies have created a revolution in user-generated information, online human networks, and rich human behavior-related data. However, the misuse of social technologies such as social media (SM) platforms, has introduced a new form of aggression and violence that occurs exclusively online. A new means of demonstrating aggressive behavior in SM websites are highlighted in this paper. The motivations for the construction of prediction models to fight aggressive behavior in SM are also outlined. We comprehensively review cyberbullying prediction models and identify the main issues related to the construction of cyberbullying prediction models in SM. This paper provides insights on the overall process for cyberbullying detection and most importantly overviews the methodology. Though data collection and feature engineering process has been elaborated, yet most of the emphasis is on feature selection algorithms and then using various machine learning algorithms for prediction of cyberbullying behaviors. Finally, the issues and challenges have been highlighted as well, which present new research directions for researchers to explore.

INDEX TERMS Big data, cyberbullying, cybercrime, human aggressive behavior, machine learning, online social network, social media, text classification.

I. INTRODUCTION

Machine or deep learning algorithms help researchers understand big data [1]. Abundant information on humans and their societies can be obtained in this big data era, but this acquisition was previously impossible [2]. One of the main sources of human-related data is social media (SM). By applying machine learning algorithms to SM data, we can exploit historical data to predict the future of a wide range of applications. Machine learning algorithms provide an opportunity to effectively predict and detect negative forms of human behavior, such as cyberbullying [3]. Big data analysis can uncover hidden knowledge through deep learning from raw data [1]. Big data analytics has improved several applications, and forecasting the future has even become possible through the combination of big data and machine learning algorithms [4].

An insightful analysis of data on human behavior and interaction to detect and restrain aggressive behavior involves multifaceted angles and aspects and the merging of theorems

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

and techniques from multidisciplinary and interdisciplinary fields. The accessibility of large-scale data produces new research questions, novel computational methods, interdisciplinary approaches, and outstanding opportunities to discover several vital inquiries quantitatively. However, using traditional methods (statistical methods) in this context is challenging in terms of scale and accuracy. These methods are commonly based on organized data on human behavior and small-scale human networks (traditional social networks). Applying these methods to large online social networks (OSNs) in terms of scale and extent causes several issues. On the one hand, the explosive growth of OSNs enhances and disseminates aggressive forms of behavior by providing platforms and networks to commit and propagate such behavior. On the other hand, OSNs offer important data for exploring human behavior and interaction at a large scale, and these data can be used by researchers to develop effective methods of detecting and restraining misbehavior and/or aggressive behavior. OSNs provide criminals with tools to perform aggressive actions and networks to commit misconduct. Therefore, methods that address both aspects (content and network) should be optimized to detect and restrain aggressive behavior in complex systems.

The remainder of this paper is organized as follows. Subsection I.A presents an overview of aggressive behavior in SM, and a new means in which SM websites are utilized by users to commit aggressive behavior is highlighted. I.B summarizes the motivations for constructing prediction models to combat aggressive behavior in SM. I.C highlight the importance of constructing cyberbullying prediction models. I.D, provide the methodology followed in this paper. Section 2 presents a comprehensive review of cyberbullying prediction models for SM websites from data collection to evaluation. Section 3 discusses the main issues related to the construction of cyberbullying prediction models. Research challenges, which present new research directions, are discussed in Section 4, and the paper is concluded in Section 5.

A. RISE OF AGGRESSIVE BEHAVIOR ON SM

Prior to the innovation of communication technologies, social interaction evolved within small cultural boundaries, such as locations and families [5]. The recent development of communication technologies exceptionally transcends the temporal and spatial limitations of traditional communication. In the last few years, online communication has shifted toward user-driven technologies, such as SM websites, blogs, online virtual communities, and online sharing platforms. New forms of aggression and violence emerge exclusively online [6]. The dramatic increase in negative human behavior on SM, with high increments in aggressive behavior, presents a new challenge [6], [7]. The advent of Web 2.0 technologies, including SM websites that are often accessed through mobile devices, has completely transformed functionality on the side of users [8]. SM characteristics, such as accessibility, flexibility, being free, and having well-connected social networks, provide users with liberty and flexibility to post and write on their platforms. Therefore, users can easily demonstrate aggressive behavior [9], [10]. SM websites have become dynamic social communication websites for millions of users worldwide. Data in the form of ideas, opinions, preferences, views, and discussions are spread among users rapidly through online social communication. The online interactions of SM users generate a huge volume of data that can be utilized to study human behavioral patterns [11]. SM websites also provide an exceptional opportunity to analyze patterns of social interactions among populations at a scale that is much larger than before.

Aside from renovating the means through which people are influenced, SM websites provide a place for a severe form of misbehavior among users. Online complex networks, such as SM websites, changed substantially in the last decade, and this change was stimulated by the popularity of online communication through SM websites. Online communication has become an entertainment tool, rather than serving only to communicate and interact with known and unknown users. Although SM websites provide many benefits to users, cyber criminals can use these websites to commit different types of misbehavior and/or aggressive behavior. The common forms of misbehavior and/or aggressive behavior on OSN sites include cyberbullying [3], phishing [12], spam distribution [13], malware spreading [14], and cyberbullying [15].

Users utilize SM websites to demonstrate different types of aggressive behavior. The main involvement of SM websites in aggressive behavior can be summarized in two points [9], [15].

1) [I.] OSN communication is a revolutionary trend that exploits Web 2.0. Web 2.0 has new features that allow users to create profiles and pages, which, in turn, make users active. Unlike Web 1.0 that limits users to being passive readers of content only, Web 2.0 has expanded capabilities that allow users to be active as they post and write their thoughts. SM websites have four particular features, namely, collaboration, participation, empowerment, and timeliness [16]. These characteristics enable criminals to use SM websites as a platform to commit aggressive behavior without confronting victims [9], [15]. Examples of aggressive behavior are committing cyberbullying [17]–[19] and financial fraud [20], using malicious applications [21], and implementing social engineering and phishing [12].
2) [II.] SM websites are structures that enable information exchange and dissemination. They allow users to effortlessly share information, such as messages, links, photos, and videos [22]. However, because SM websites connect billions of users, they have become delivery mechanisms for different forms of aggressive behavior at an extraordinary scale. SM websites help cybercriminals reach many users [23].

B. MOTIVATIONS FOR PREDICTING AGGRESSIVE BEHAVIOR ON SM WEBSITES

Many studies have been conducted on the contribution of machine learning algorithms to OSN content analysis in the last few years. Machine learning research has become crucial in numerous areas and successfully produced many models, tools, and algorithms for handling large amounts of data to solve real-world problems [24], [25]. Machine learning algorithms have been used extensively to analyze SM website content for spam [26]–[28], phishing [29], and cyberbullying prediction [19], [30]. Aggressive behavior includes spam propagation [13], [31]–[34], phishing [12], malware spread [14], and cyberbullying [15]. Textual cyberbullying has become the dominant aggressive behavior in SM websites because these websites give users full freedom to post on their platforms [17], [35]–[39].

SM websites contain large amounts of text and/or non-text content and other information related to aggressive behavior. In this work, a content analysis of SM websites is performed to predict aggressive behavior. Such an analysis is limited to textual OSN content for predicting cyberbullying behavior. Given that cyberbullying can be easily committed, it is considered a dangerous and fast-spreading aggressive behavior. Bullies only require willingness and a laptop or cell phone with Internet connection to perform misbehavior without confronting victims [40]. The popularity and proliferation of SM websites have increased online bullying activities. Cyberbullying in SM websites is rampant due to the structural characteristics of SM websites. Cyberbullying in traditional platforms, such as emails or phone text messages, is performed on a limited number of people. SM websites allow users to create profiles for establishing friendships and communicating with other users regardless of geographic location, thus expanding cyberbullying beyond physical location. Anonymous users may also exist on SM websites, and this has been confirmed to be a primary cause for increased aggressive user behavior [41]. Developing an effective prediction model for predicting cyberbullying is therefore of practical significance. With all these considerations, this work performs a content-based analysis for predicting textual cyberbullying on SM websites.

The motivation of this review is explained in the following section.

C. WHY CONSTRUCTING CYBERBULLYING PREDICTION MODELS IS IMPORTANT

The motivations for carrying out this review for predicting cyberbullying on SM websites are discussed as follows. Cyberbullying is a major problem [42] and has been documented as a serious national health problem [43] due to the recent growth of online communication and SM websites. Research has shown that cyberbullying exerts negative effects on the psychological and physical health and academic performance of people [44]. Studies have also shown that cyberbullying victims incur a high risk of suicidal ideation [45], [46]. Other studies [45], [46] reported an association between cyberbullying victimization and suicidal ideation risk. Consequently, developing a cyberbullying prediction model that detects aggressive behavior that is related to the security of human beings is more important than developing a prediction model for aggressive behavior related to the security of machines.

Cyberbullying can be committed anywhere and anytime. Escaping from cyberbullying is difficult because cyberbullying can reach victims anywhere and anytime. It can be committed by posting comments and statuses for a large potential audience. The victims cannot stop the spread of such activities [47]. Although SM websites have become an integral part of users' lives, a study found that SM websites are the most common platforms for cyberbullying victimization [48]. A well-known characteristic of SM websites, such as Twitter, is that they allow users to publicly express and spread their posts to a large audience while remaining anonymous [9]. The effects of public cyberbullying are worse than those of private ones, and anonymous scenarios of cyberbullying are worse than non-anonymous cases [49], [50]. Consequently, the severity of cyberbullying has increased on SM websites, which support public and anonymous scenarios of cyberbullying. These characteristics make SM websites, such as Twitter, a dangerous platform for committing cyberbullying [43].

Recent research has indicated that most experts favor the automatic monitoring of cyberbullying [51]. A study that examined 14 groups of adolescents confirmed the urgent need for automatic monitoring and prediction models for cyberbullying [52] because traditional strategies for coping with cyberbullying in the era of big data and networks do not work well. Moreover, analyzing large amounts of complex data requires machine learning-based automatic monitoring.

1) CYBERBULLYING ON SM WEBSITES

Most researchers define cyberbullying as using electronic communication technologies to bully people [53]. Cyberbullying may exist in different types or forms, such as writing aggressive posts, harassing or bullying a victim, making hateful posts, or insulting the victim [54], [55]. Given that cyberbullying can be easily committed, it is considered a dangerous and fast-spreading aggressive behavior. Bullies only require willingness and a laptop or cell phone connected to the Internet to perform misbehavior without confronting the victims [40]. The popularity and proliferation of SM websites have increased online bullying activities. Cyberbullying on SM websites is performed on a large number of users due to the structural characteristics of SM websites [48].

Cyberbullying in traditional platforms, such as emails or phone text messages, is committed on a limited number of people. SM websites allow users to create profiles for establishing friendships and interacting with other online users regardless of geographic location, thus expanding cyberbullying beyond physical location. Moreover, anonymous users may exist on SM websites, and this has been confirmed to be a primary cause of increased aggressive user behavior [41].

The nature of SM websites allows cyberbullying to occur secretly, spread rapidly, and continue easily [54]. Consequently, developing an effective prediction model for predicting cyberbullying is of practical significance. SM websites contain large amounts of text and/or non-text content and information related to aggressive behavior.

D. METHODOLOGY

This section presents the methodology used in this work for a literature search. Two phases were employed to retrieve published papers on cyberbullying prediction models. The first phase included searching for reputable academic databases and search engines. The search engines and academic databases used for the retrieval of relevant papers were as follows: Scopus, Clarivate Analytics' Web of Science, DBLP Computer Science Bibliography, ACM Digital Library, ScienceDirect, SpringerLink, and IEEE Xplore. The major keywords used for the literature search were coined in relation to social media as follows: cyberbullying, aggressive behavior, big data, and cyberbullying models. The second phase involved searching for literature through Qatar University's digital library. The articles retrieved from the search were scrutinized to ensure that the articles met the inclusion criteria. According to the inclusion criteria, for an article to be selected for the survey, it must report an empirical study describing the prediction of cyberbullying on SM sites. Otherwise, the article would be excluded in the selection. Many articles were rejected based on titles. The abstract and conclusion sections were examined to ensure that articles satisfied the screening criteria, and those that did not satisfy the criteria were excluded from the survey.

II. PREDICTING CYBERBULLYING ON SOCIAL MEDIA IN THE BIG DATA ERA USING MACHINE LEARNING ALGORITHMS

Our world is currently in the big data era because 2.5 quintillion bytes of data are generated daily [56]. Organizations continuously generate large-scale data. These large-scale datasets are generated from different sources, including the World Wide Web, social networks, and sensor networks [57]. Big data have nine characteristics, namely, volume, variety, variability and complexity, velocity, veracity, value, validity, verdict, and visibility [58]. For example, Flickr generates almost 3.6 TB of data, Google is believed to process almost 20,000 TB of data per day, and the Internet gathers an estimated 1.8 PB of data daily [59].

SM is an online platform that provides users an opportunity to create an online community, share information, and exchange content. SM users and the interaction among organizations, people, and products are responsible for the massive amount of data generated on SM platforms. SM platforms, such as Facebook, YouTube, blogs, Instagram, Wikipedia, and Twitter, are of different types. The data generated by SM outlets can be structured or unstructured in form. SM analytics is the analysis of structured and unstructured data generated by SM outlets. SM analytics can be in any of the following forms: link prediction, community, content, social influence, structured, and unstructured. SM is now in the big data era. For example, Facebook stores 260 billion photographs in over 20 PB of storage space, and up to one million pictures are processed per second. YouTube receives 100 hours of downloaded videos in each minute [60].

The most common means of constructing cyberbullying prediction models is to use a text classification approach that involves the construction of machine learning classifiers from labeled text instances [19], [38], [61]–[63]. Another means is to use a lexicon-based model that involves computing orientation for a document from the semantic orientation of words or phrases in the document [64]. Generally, the lexicon in lexicon-based models can be constructed manually (similar to the approaches used in [65]) or automatically by using seed words to expand the list of words [66]. However, cyberbullying prediction using the lexicon-based approach is rare in literature. The primary reason is that the texts on SM websites are written in an unstructured manner, thus making it difficult for the lexicon-based approach to detect cyberbullying based only on lexicons [67]–[69]. However, lexicons are used to extract features, which are often utilized as inputs to machine learning algorithms. For example, lexicon-based approaches, such as using a profane-based dictionary to detect the number of profane words in a post, are adopted as profane features to machine learning models [70]. The key to effective cyberbullying prediction is to have a set of features that are extracted and engineered [71]. Features and their combinations are crucial in the construction of effective cyberbullying prediction models [70], [71]. Most studies on cyberbullying prediction [19], [38], [62], [72], [73] used machine learning algorithms to construct cyberbullying prediction models. Machine learning-based models exhibit decent performance in cyberbullying prediction [74]. Consequently, this work reviews the construction of cyberbullying prediction models based on machine learning.

The machine learning field focuses on the development and application of computer algorithms that improve with experience [75], [76]. The objective of machine learning is to identify and define the patterns and correlations between data. The importance of analyzing big data lies in discovering hidden knowledge through deep learning from raw data [1]. Machine learning can be described as the adoption of computational models to improve machine performance by predicting and describing meaningful patterns in training data and the acquisition of knowledge from experience [77]. When this concept is applied to OSN content, the potential of machine learning lies in exploiting historical data to detect, predict, and understand large amounts of OSN data. For example, in supervised machine learning for classification application, classification is learned with the help of suitable examples from a training dataset. In the testing stage, new data are fed into the model, and instances are classified to a specified class learned during the training stage. Then, classification performance is evaluated.

This section reviews the most common processes in the construction of cyberbullying prediction models for SM websites based on machine learning. The review covers data collection, feature engineering, feature selection, and machine learning algorithms.

A. DATA COLLECTION

Data are important components of all machine learning-based prediction models. However, data (even ''Big Data'') are useless on their own until knowledge or implications are extracted from them. Data extracted from SM websites are used to select training and testing datasets. Supervised prediction models aim to provide computer techniques to enhance prediction performance in defined tasks on the basis of observed instances (labeled data) [78]. Machine learning models for a certain task primarily aim to generalize; a successful model should not be limited to examples in a training dataset only [79] but must include unlabeled real data. Data quantity is inconsequential; what is crucial is whether or not the extracted data represent activities on SM websites well [80]–[82]. The main data collection strategies in previous cyberbullying prediction studies on SM websites can be categorized into data extracted from SM websites by using either keywords, that is, words, phrases, or hashtags (e.g., [19], [43], [83]–[85]), or by using user profiles (e.g., [38], [62], [70], [86]). The issues in these data collection strategies and their effects on the performance of machine learning algorithms are highlighted in the Data Collection section (related issues).

B. FEATURE ENGINEERING

Feature is a measurable property of a task that is being observed [87]. The main purpose of engineering feature vectors is to provide machine learning algorithms with a set of learning vectors through which these algorithms learn how to discriminate between different types of classes [76]. Feature engineering is a key factor behind the success and failure of most machine learning models [79]. The success and failure of prediction may be based on several elements. The most significant element is the features used to train the model [78]. Most of the effort in constructing cyberbullying prediction models using learning algorithms is devoted to this task [61], [62], [72]. In this context, the design of the input space (i.e., features and their combinations that are provided as an input to the classifier) is vital.

Proposing a set of discriminative features, which are used as inputs to the machine learning classifier, is the main step toward constructing an effective classifier in many applications [76]. Feature sets can be created based on human-engineered observations, which rely on how features correlate with the occurrences of classes [76]. For example, recent cyberbullying studies [88]–[94] established the correlation between different variables, such as age, gender, and user personality, and cyberbullying occurrence. These observations can be engineered into a practical form (feature) to allow the classifier to discriminate between cyberbullying and non-cyberbullying and can thus be used to develop effective cyberbullying prediction models. Proposing features is an important step toward improving the discrimination power of prediction models [76], [79]. Similarly, proposing a set of significant features of cyberbullying engagement on SM websites is important in developing effective prediction models based on machine learning algorithms [68], [95].

State-of-the-art research has developed features to improve the performance of cyberbullying prediction. For example, a lexical syntactic feature has been proposed to deal with the prediction of offensive language; this method is better than traditional learning-based approaches in terms of precision [18]. Dadvar *et al.* examined gender information from profile information and developed a gender-based approach for cyberbullying prediction by using datasets from Myspace as a basis. The gender feature was selected to improve the discrimination capability of a classifier. Age and gender were included as features in other studies [17], [61], but these features are limited to the information provided by users in their online profiles.

Several studies focused on cyberbullying prediction based on profane words as a feature [35], [68], [70], [95], [96]. Similarly, a lexicon of profane words was constructed to indicate bullying, and these words were used as features for input to machine learning algorithms [97], [98]. Using profane words as features demonstrates a significant improvement in model performance. For example, the number of ''bad'' words and the density of ''bad'' words were proposed as features for input to machine learning in a previous work [70]. The study concluded that the percentage of ''bad'' words in a text is indicative of cyberbullying. Another research [85] expanded a list of pre-defined profane words and allocated different weights to create bullying features. These features were concatenated with bag-of-words and latent semantic features and used as a feature input for a machine learning algorithm.

Reference [19] proposed features, such as pronouns and skip grams, as additional features to traditional models, such as bag of words (n-gram n = 1). The authors claimed that adding these features improved the overall classification accuracy. Another study [62] analyzed textual cyberbullying associated with comments on images in Instagram and developed a set of features from text comprising traditional bag-of-words features, comment counts for an image, and post counts within less than one hour of posting the image. Features mined from user and media information, including the number of followers and likes, and shared media and features from image content, such as image types, were added [62]. The combination of all features improved the overall classification performance [62].

The context-based approach is better than the list-based approach in developing the feature vector [37]. However, the diversity and complexity of cyberbullying do not always support this conclusion. Several studies [68], [72], [96], [99] discussed how sentiment analysis can improve the discrimination power of a classifier to distinguish between

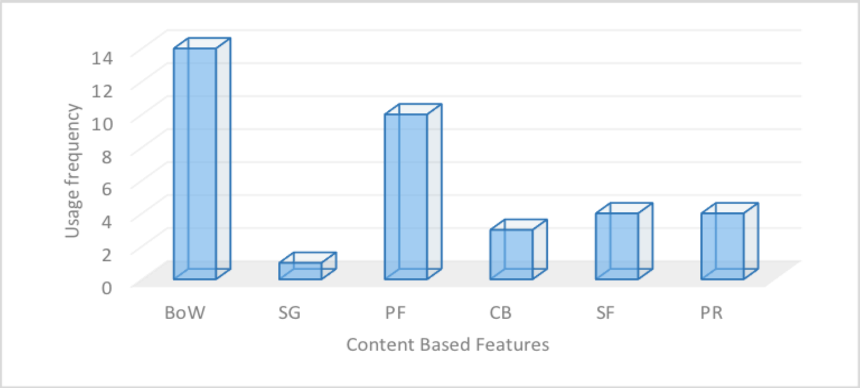


FIGURE 1. Depicting feature types used in cyberbullying prediction: Content-based features.

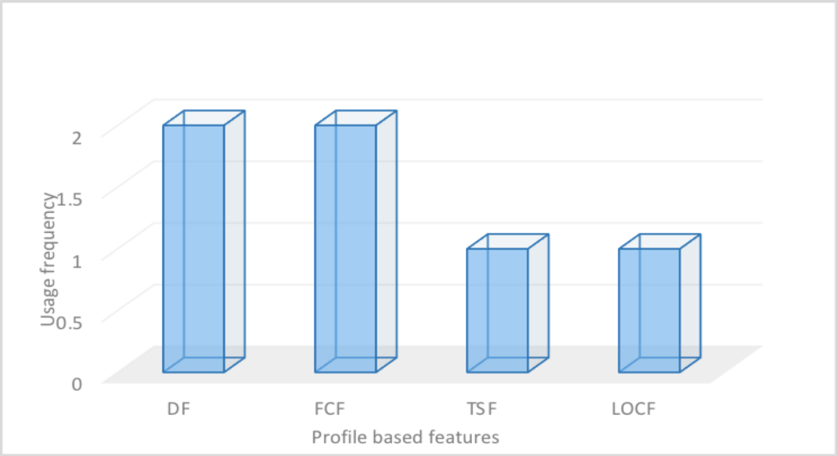


FIGURE 2. Depicting feature types used in cyberbullying prediction: Profile-based features.

cyberbullying and normal posts. These studies assumed that sentiment features are a good signal for cyberbullying occurrence. In another study that aimed to establish ways of reducing cyberbullying activities by predicting troll profiles, the researchers proposed a model to identify and associate troll profiles in Twitter; they assumed that predicting troll profiles is an important step toward predicting and stopping cyberbullying occurrence on SM websites [38]. This study proposed features based on tweeted text, posting time, language, and location to improve the identification of authorship of posts and determine whether a profile is troll or not. Reference [99] merged features from the structure of SM websites (e.g., degree, closeness, betweenness, and eigenvector centralities as well as clustering coefficient) with features from users (e.g., age and gender) and content (e.g., length and sentiment of a post). Combining these features improves the final machine learning accuracy [99]. Table 1 shows a comparison of the different features used in cyberbullying prediction literature. affect prediction performance. If the constructed features contain a large set of features that individually associate well with class, then the learning process will be effective. This condition explains why most of the discussed studies aimed to produce many features. The input features should reflect the behavior related to the occurrence of textual cyberbullying. However, the set of features should be analyzed using feature selection algorithms. Feature selection algorithms are adopted to decide which features are most probably relevant or irrelevant to classes.

C. FEATURE SELECTION ALGORITHMS

Feature selection algorithms were rarely adopted in state-of-the-art research to perform cyberbullying prediction on SM websites via machine learning (all extracted features are used to train the classifiers). Most of the examined studies (e.g., [18], [61], [68], [70]–[72], [85], [95], [96], [99]) did not use feature selection to decide which features are important in training machine learning algorithms. Two studies [19], [62] used chi-square and PCA to select a significant feature from extracted features. These feature selection algorithms are briefly discussed in following subsections.

1) INFORMATION GAIN

Information gain is the estimated decrease in entropy produced by separating examples based on specified features. Entropy is a well-known concept in information theory; it describes the (im)purity of an arbitrary collection of examples [100].

TABLE 1. Summary of feature types used in cyberbullying prediction literature.

Study	Content-based Features						Profile-based Features			
	BoW	SG	PF	CB	SF	PR	DF	FCF	TSF	LOCF
[19]	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗
[18]	✓	✗	✓	✓	✗	✓	✗	✗	✗	✗
[61]	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗
[95]	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗
[72]	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗
[62]	✓	✗	✓	✗	✗	✗	✗	✓	✗	✗
[68]	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗
[74]	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
[85]	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗
[99]	✓	✗	✓	✗	✓	✓	✓	✓	✗	✗
[70]	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗
[96]	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗
[43]	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
[38]	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓
[71]	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗

**BoW = bag of words, SG = skip gram, PF = profanity features, SF = sentiment features, PR = pronouns, DF = demographic features (e.g., age and gender), FCF = friends or follower count features, TSF = timestamp features, LOCF = location of post feature

Information gain is used to calculate the strength or importance of features in a classification model according to the class attribute. Information gain [101] evaluates how well a specified feature divides training datasets with respect to class labels, as explained in the following equations. Given a training dataset (Tr), the entropy of (Tr) is defined as.

$$I(Tr) = -\sum P_n \log_2 P_n, \tag{1}$$

where P_n is the probability that Tr belongs to class n.

For attribute Att datasets, the expected entropy is calculated as

$$I(Att) = \sum \left(\frac{Tr_{Att}}{Tr}\right) \times I\left(Tr_{Att}\right). \tag{2}$$

The information gain of attribute Att datasets is

$$IG(Att) = I(Tr) - I(Att) \tag{3}$$

2) PEARSON CORRELATION

Correlation-based feature selection is commonly used in reducing feature dimensionality and evaluating the discrimination power of a feature in classification models. It is also a straightforward model for selecting significant features. Pearson correlation measures the relevance of a feature by computing the Pearson correlation between it and a class. The Pearson correlation coefficient measures the linear correlation between two attributes [102]. The subsequent value lies between -1 and $+1$, with -1 implying absolute negative correlation (as one attribute increases, the other decreases), $+1$ denoting absolute positive correlation (as one attribute increases, the other also increases), and 0 denoting the absence of any linear correlation between the two attributes. For two attributes or features X and Y, the Pearson correlation coefficient measures the correlation [103] as follows:

$$r_{xy} = \frac{\sum (x_i - \overline{x})\,(y_i - \overline{y})}{(n-1)\,S_x S_y}, \tag{4}$$

where $\overline{x}$ and $\overline{y}$ are the sample means for X and Y, respectively; S_x and S_y are the sample standard deviations for X and Y, respectively; and n is the size of the sample used to compute the correlation coefficient [103].

3) CHI-SQUARE TEST

Another common feature selection model is the chi-square test. This test is used in statistics, among other variables, to test the independence of two occurrences. In feature selection, chi-square is used to test whether the occurrences of a feature and class are independent. Thus, the following quantity is assumed for each feature, and they are ranked by their score.

$$N = \frac{N\left[P(f, c_i)P(\overline{f}, \overline{c}_i) - P(f, \overline{c}_i)P(\overline{f}, c_i)\right]}{P(f)P(\overline{f})P(c_i)P(\overline{c}_i)} \tag{5}$$

The chi-square test [104] assesses the independence between feature f and class c_i, in which N is the total number of documents.

D. MACHINE LEARNING ALGORITHMS

Many types of machine learning algorithms exist, but nearly all studies on cyberbullying prediction in SM websites used the most established and widely used type, that is, supervised machine learning algorithms [67], [99]. The accomplishment of machine learning algorithms is determined by the degree to which the model accurately converts various types of prior observation or knowledge about the task. Much of the practical application of machine learning considers the details

TABLE 2. Summary of machine learning algorithms tested in cyberbullying literature.

Study	SVM	NB	RF	DT	KNN	LR	ARM	RB
[19]	✓	✗	✗	✗	✗	✓	✗	✗
[18]	✓	✓	✗	✗	✗	✗	✗	✗
[61]	✓	✗	✗	✗	✗	✗	✗	✗
[95]	✓	✓	✗	✓	✗	✗	✗	✓
[38]	✓	✓	✓	✓	✓	✗	✗	✗
[86]	✗	✗	✓	✗	✗	✗	✗	✗
[72]	✓	✗	✗	✗	✗	✗	✗	✗
[62]	✓	✗	✗	✗	✗	✗	✗	✗
[74]	✗	✓	✗	✗	✗	✗	✗	✗
[73]	✓	✓	✗	✗	✗	✓	✗	✗
[84]	✗	✗	✗	✗	✗	✗	✓	✗
[71]	✓	✗	✗	✗	✗	✗	✗	✗
[85]	✓	✗	✗	✗	✗	✗	✗	✗
[99]	✓	✓	✗	✗	✗	✓	✗	✗
[70]	✓	✗	✗	✓	✗	✗	✗	✓
[96]	✓	✗	✗	✗	✗	✗	✗	✗

** SVM = support vector machine family, NB = naïve Bayes, RF = random forest, DT = decision tree family, KNN = K-nearest neighbor, LR = logistic regression, ARM = association rule mining, RB = rule-based algorithms

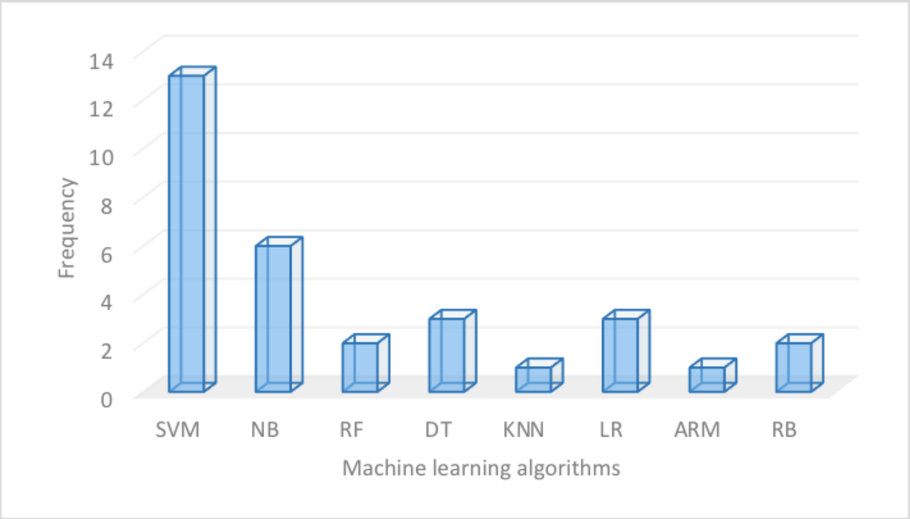


FIGURE 3. Machine learning algorithms applied in cyberbullying prediction.

of a particular problem. Then, an algorithmic model that allows for the accurate encoding of the facts is selected. However, no optimal machine learning algorithm works best for all problems [73], [105], [106]. Therefore, most researchers selected and compared many supervised classifiers to determine the ideal ones for their problem. Classifier selection is generally based on the most commonly used classifiers in the field and the data features available for experiments. However, researchers can only decide which algorithms to adopt for constructing a cyberbullying prediction model by performing a comprehensive practical experiment as a basis. Table 2 summarizes the commonly used machine learning algorithms for constructing cyberbullying prediction models.

The following sections describe the machine learning algorithms commonly used for constructing cyberbullying prediction models (Table 2).

1) SUPPORT VECTOR MACHINE IN CYBERBULLYING

Support vector machine (SVM) is a supervised machine learning classifier that is commonly used in text classification [107]. SVM is constructed by generating a separating hyperplane in the feature attributes of two classes, in which the distance between the hyperplane and the adjacent data point of each class is maximized [108]. Theoretically, SVM was developed from statistical learning theory [109]. In the SVM algorithm, the optimal separation hyperplane pertains to the separating hyperplane that minimizes misclassifications that is achieved in the training step. The approach is based on minimized classification risks [106], [110]. SVM was initially established to classify linearly separable classes. A 2D plane comprises linearly separable objects from different classes (e.g., positive or negative). SVM aims to separate the two classes effectively. SVM identifies the exceptional hyperplane that provides the maximum margin by maximizing the distance between the hyperplane and the nearest data point of each class.

In real-time applications, precisely determining the separating hyperplane is difficult and nearly impossible in several cases. SVM was developed to adapt to these cases and can now be used as a classifier for non-separable classes. SVM is a capable classification algorithm because of its characteristics. Specifically, SVM can powerfully separate non-linearly divisible features by converting them to a high-dimensional space using the kernel model [111].

The advantage of SVM is its high speed, scalability, capability to predict intrusions in real time, and update training patterns dynamically.

SVM has been used to develop cyberbullying prediction models and found to be effective and efficient. For example, Chen *et al.* [18] applied SVM to construct a cyberbullying prediction model for the detection of offensive content in SM. SM content with potential cyberbullying were extracted, and the SVM cyberbullying prediction model was applied to detect offensive content. The result showed that SVM is more accurate in detecting user offensiveness than naïve Bayes (NB). However, NB is faster than SVM. Chavan and Shylaja [19] proposed the use of SVM to build a classifier for the detection of cyberbullying in social networking sites. Data containing offensive words were extracted from social networking sites and utilized to build a cyberbullying SVM prediction model. The SVM classifier detected cyberbullying more accurately than LR did. Dadvar *et al.* [61] used SVM to build a gender specific cyberbullying prediction model. An SVM text classifier was created with gender specific characteristics.

The SVM cyberbullying prediction model enhanced the detection of cyberbullying in SM. Hee *et al.* [72] developed an SVM-based cyberbullying detection model to detect cyberbullying in a social network site. The SVM-based model was trained using data containing cyberbullying extracted from the social network site. The researchers found that that the SVM-based cyberbullying model effectively detected cyberbullying. Mangaonkar *et al.* [73] constructed an SVM-based cyberbullying detection model for YouTube. Data were collected from YouTube comments on videos posted on the site. The data were used to train SVM and construct a cyberbullying detection model, which was then used to detect cyberbullying. The results suggested that the SVM-based cyberbullying model is more reliable but not as accurate as rule-based Jrip. However, the SVM-based cyberbullying model is more accurate than NB and tree-based J48. Dinakar *et al.* [95] proposed the use of SVM for the detection of cyberbullying in Twitter. An SVM-based cyberbullying model was constructed from data extracted from Twitter. The SVM-based cyberbullying prediction model was applied to detect cyberbullying in Twitter. SVM detected cyberbullying better than NB- and LR-based cyberbullying detection models did.

2) NB ALGORITHM

NB was used to construct cyberbullying prediction models in [18], [38], [73], [74], and [95]. NB classifiers were constructed by applying Bayes' theorem between features. Bayesian learning is commonly used for text classification. This model assumes that the text is generated by a parametric model and utilizes training data to compute Bayes-optimal estimates of the model parameters. It categorizes generated test data with these approximations [112].

NB classifiers can deal with an arbitrary number of continuous or categorical independent features [106]. By using the assumption that the features are independent, a high-dimensional density estimation task is reduced to one-dimensional kernel density estimation [106].

The NB algorithm is a learning algorithm that is grounded on the use of Bayes theorem with strong (naive) independence assumptions. This method was discussed in detail in [113]. The NB algorithm is one of the most commonly used machine learning algorithms [114], and it has been constructed as a machine learning classifier in numerous social media based studies [115]–[117].

3) RANDOM FOREST

Random forest (RF) was used in the construction of cyberbullying prediction models in [72] and [86]. RF is a machine-learning model that combines decision trees and ensemble learning [118]. This model fits several classification trees to a dataset then combines the predictions from all the trees [119]. Therefore, RF consists of many trees that are used randomly to select feature variables for the classifier input. The construction of RF is achieved in the following simplified steps.

1. The number of examples (cases) in training data is set to N, and the number of attributes in the classifier is M.
2. A number of random decision tress is created by selecting attributes randomly. A training set is selected for each tree by choosing n times from all N existing instances. The rest of the instances in the training set are used to approximate the error of the tree by forecasting their classes.
3. For each tree's nodes, m random variables are selected on which to base the decision at that node. The finest split is computed using these m attributes in the training set. Each tree is completely built and is not pruned, as can be done in building a normal tree classifier.
4. A large number of trees are thus created. These decision trees vote for the most popular class. These processes are called RFs [118].

RF constructs a model that comprises a group of tree-structured classifiers, in which each tree votes for the most popular class [118]. The most highly voted class is the selected as the output.

4) DECISION TREE

Decision tree classifiers were used in construction of cyberbullying prediction models in [38] and [95]. Decision trees are easy to understand and interpret; hence, the decision tree algorithm can be used to analyze data and build a graphic model for classification. The most commonly improved version of decision tree algorithms used for cyberbullying prediction is C.45 [38], [70], [95]. C4.5 can be explained as follows. Given N number of examples, C4.5 first produces an initial tree through the divide-and-conquer algorithm as follows [120]:

If all examples in N belong to the same class or N is small, the tree is a leaf labeled with the most frequent class in N. Otherwise, a test is selected based on, for example, the mostly used information gain test on a single attribute with two or more outputs. Considering that the test is the root of the tree creation partition of N into subsets $N_1, N_2, N_3 \ldots\ldots.$ regarding the outputs for each example, the same procedure is applied recursively to each subset [120].

5) K-NEAREST NEIGHBOR

K-nearest neighbor (KNN) is a nonparametric technique that decides the KNNs of X_0 and uses a majority vote to calculate the class label of X_0. The KNN classifier often uses Euclidean distances as the distance metric [121]. To demonstrate a KNN classification, classifying new input posts (from a testing set) is considered by using a number of known manually labeled posts. The main task of KNN is to classify the unknown example based on a nominated number of its nearest neighbors, that is, to finalize the class of unknown examples as either a positive or negative class. KNN classifies the class of unknown examples by using majority votes for the nearest neighbors of the unknown classes. For example, if KNN is one nearest neighbor [estimating the class of an unknown example using the one nearest neighbor vote (k = 1)], then KNN will classify the class of the unknown example as positive (because the closest point is positive). For two nearest neighbors (estimating the class of an unknown example using the two nearest neighbor vote), KNN is unable to classify the class of the unknown example because the second closest point is negative (positive and negative votes are equal). For four nearest neighbors (estimating the class of an unknown example using the four nearest neighbor vote), KNN classifies the class of the unknown example as positive (because the three closest points are positive and only one vote is negative). The KNN algorithm is one of the simplest classification algorithms, but despite its simplicity, it can provide competitive results [122]. KNN was used in the construction of cyberbullying prediction models in [38].

6) LOGISTIC REGRESSION CLASSIFICATION

Logistic regression is one of the common techniques imported by machine learning from the statistics field. Logistic regression is an algorithm that builds a separating hyperplane between two datasets by means of the logistic function [123]. The logistic regression algorithm takes inputs (features) and generates a forecast according to the probability of the input being appropriate for a class. For example, if the probability is >0.5, the classification of the instance will be a positive class; otherwise, the prediction is for the other class (negative class) [124]. Logistic regression was used in the construction of cyberbullying prediction models in [19] and [73].

E. EVALUATION

The primary objective of constructing prediction models based on machine learning is to generalize more than the training dataset [79]. When a machine learning model is applied to a real example, it can perform well. Accordingly, the data are divided into two parts. The first part is the training data used to train machine learning algorithms. The second part is the testing data used to test machine learning algorithms. However, separately dividing data into training and testing is not widely employed [79], especially in applications in which deriving training and testing data are difficult. For example, in cyberbullying prediction, most state-of-art studies manually labeled data. Hence, creating labeled data is expensive. These issues can be reduced by cross validation, that is, randomly dividing the training data into 10 subsets for example, and this process is called 10-fold cross validation. Cross validation involves the following steps: keep a fold separate (the model does not see it) and train data on the model by using the remaining folds; test each learned classifier on the fold which it did not see; and average the results to see how well the particular parameter setting performs [79], [125].

F. EVALUATION METRICS

Researchers measure the effectiveness of a proposed model to determine how successfully the model can distinguish cyberbullying from non-cyberbullying by using various evaluation measures. Reviewing common evaluation metrics in the research community is important to understand the performance of conflicting models. The most commonly used metrics in evaluating cyberbullying classifiers for SM websites are as follows:

1) ACCURACY

It was used to evaluate cyberbullying prediction models in [62], [70], [73] and [95], and it is calculated as follows:

$$Accuracy = \frac{(tp + tn)}{(tp + fp + tn + fn)}. \tag{6}$$

2) PRECISION, RECALL, AND F-MEASURE

These were used to evaluate cyberbullying prediction models in [18], [61], [72], and [73]. They are calculated as follows:

$$Precision = \frac{tp}{(tp + fp)}, \tag{7}$$

$$Recall = \frac{tp}{(tp + fn)}, \tag{8}$$

$$F - Measure = \frac{2 \times precision \times recall}{recision + recall} \tag{9}$$

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

3) AREA UNDER THE CURVE (AUC)

AUC offers a discriminatory rate of the classifier at various operating points [3], [19], [38]. The main benefit of using AUC as an evaluation metric is that AUC gives a more robust measurement than the accuracy metric in class-imbalance situations [19], [38].

III. ISSUES RELATED TO CONSTRUCTING CYBERBULLYING PREDICTION MODELS

In this section, the issues identified from the reviewed studies are discussed. The main issues related to cyberbullying definition, data collection feature engineering, and evaluation metric selection are identified and discussed in following subsections.

A. ISSUES RELATED TO CYBERBULLYING DEFINITION

Traditional bullying is generally defined as "intentional behavior to harm another, repeatedly, where it is difficult for the victim to defend himself or herself" [126]. By extending the definition of traditional bullying, cyberbullying has been defined [90] as "an aggressive behavior that is achieved using electronic platforms by a group or an individual repeatedly and over time against a victim who cannot easily defend him or herself." Applying such a definition makes it difficult to classify manually labeled data (the instance in which machine learning algorithms learn from) and whether a post is cyberbullying or not. Two main issues make the above definition difficult to be applied in online environments [47], [127]. The first issue is how to measure "repeatedly and over time aggressive behavior" on SM, and the second one is how to measure power imbalance and "a victim who cannot easily defend himself or herself" on SM. These issues have been discussed by researchers to simplify the concept of cyberbullying in the online context. First, the concept of repetitive act in cyberbullying is not as straightforward as that in SM [47]. For example, SM websites can provide cyberbullies a medium to propagate cyberbullying posts for a large population. Consequently, a single act by one committer may become repetitive over time [47]. Second, power imbalance is presented in different forms in online communication. Researchers [127] have suggested that the content in online environments is difficult to eliminate or avoid, thus making a victim powerless.

These definitional aspects are under intense debate, but to simplify the definition of cyberbullying and make this definition applicable to a wide range of applications, the researchers in [53] and [72] defined cyberbullying as "the use of electronic communication technologies to bully others." Proposing a simplified and clear definition of cyberbullying is a crucial step toward building machine learning models that can satisfy the definition criteria of cyberbullying engagement.

B. DATA COLLECTION

Many cyberbullying prediction studies extracted their datasets by using specific keywords or profile IDs. Nevertheless, by simply tracking posts that have particular keywords, these researches may have presented potential sampling bias [82], [128], limited the prediction to posts that contain the predefined keywords, and overlooked many other posts relevant to cyberbullying. Such data collection methods limit the prediction model of cyberbullying to specified keywords. The identification of keywords for extracting posts is also subject to the author's understanding of cyberbullying. An effective method should use a complete range of posts indicating cyberbullying to train the machine learning classifier and ensure the generalization capability of the cyberbullying prediction model [43]. An important objective of machine learning is to generalize and not to limit the examples in a training dataset [79]. Researchers should investigate whether the sampled data are extracted from data that effectively represents all possible activities on SM websites [128]. Extracting well-representative data from SM is the first step toward building effective machine learning prediction models. However, SM websites' public application program interface (API) only allows the extraction of a small sample of all relevant data and thus poses a potential for sampling bias [80]–[82]. For example, a previous study [128] discussed whether data extracted from Twitter's streaming API is a sufficient representation of the activities in the Twitter network as a whole; the author compared keyword (words, phrases, or hashtags), user ID, and geo-coded sampling. Twitter's streaming API returns a dataset with some bias when keyword or user ID sampling is used. By contrast, using geo-tagged filtering provides good data representation [128]. With these points in mind, researchers should ensure minimum bias as much as possible when they extract data to guarantee that the examples selected to be represented in training data are generalized and provide an effective model when applied to testing data. Bias in data collection can impose bias in the selected training dataset based on specific keywords or users, and such a bias consequently introduces overfitting issues that affect the capability of a machine learning model to make reliable predictions on untrained data.

C. FEATURE ENGINEERING

Features are vital components in improving the effectiveness of machine learning prediction models [79]. Most of the discussed studies attempted to provide effective machine learning solutions to cyberbullying on SM websites by providing significant features (Table 1). However, these studies overlooked other important features. For example, online cyberbullies may dynamically change the way they use words and acronyms. SM websites help create cyberbullying acronyms that have not been commonly used in committing traditional bullying or are beyond SM norms [129]. Recent survey response studies (questionnaire-based studies) have reported positive correlations between different variables, such as personality [93], [94] and sociability of a user in an online environment [130], and cyberbullying occurrences. The observations of these studies are important in understanding such behavior in online environments. However, these observations are yet to be used as features with machine learning algorithms to provide significant models. These observations can be useful when transformed to a practical form (features) that can be employed to develop effective machine learning prediction models for cyberbullying on SM websites. The abundant information provided by SM websites should be utilized to convert observations into a set of features. For example, two studies [17], [61] attempted to improve machine learning classifier performance by including features, such as age and gender, that show improvement in classifier performance, but these features are extracted from direct user details mentioned in the online profiles of users. However, most studies found that only a few users provide complete details in their online profiles [131], [132]. These studies suggested the useful practice of utilizing words expressed in the content (posts) to identify user age and gender [131], [132]. Moreover, cyberbullying is related to the aggressive behavior of a user. A study demonstrated that aggression considerably predicts cyberbullying [92]. Similarly, cyberbullying behavior has a strong correlation with neuroticism [93], [94]. Therefore, predicting if a user has used words related to neuroticism may provide a useful feature to predict cyberbullying engagement.

A significant correlation has also been found between sociability of a user and cyberbullying engagement in online environments [130]. Users who are highly active in online environments are likely to engage in cyberbullying [133]. According to these observations, SM websites possess features that can be used as signals to measure the sociability of a user, such as number of friends, number of posts, URLs in posts, hashtags in posts, and number of users engaged in conversations (mentioned). The combination of these features with traditionally used ones, such as profanity features, can provide comprehensive discriminative features. The reviewed studies (Table 1) focused on using either a traditional feature model (e.g., bag-of-words) or information (e.g., age or gender) limited to user profile information (information written by users in their profile). Given that such information is limited, comprehensive features should be proposed to improve classifier performance.

Moreover, maintaining a precise and accurate process in constructing machine learning models from start (data collection) to end (evaluation metric selection) is important in ensuring that the proposed features hold significance in improving classifier performance. The following subsection analyzes other issues related to constructing effective machine learning models for cyberbullying prediction on SM websites.

D. MACHINE LEARNING ALGORITHM SELECTION

A machine learning algorithm is selected to be trained on proposed features. However, deciding which classifier performs best for a specific dataset is difficult. More than one machine learning algorithm should be tested to determine the best machine learning algorithm for a specific dataset. Three points may be used as guide to narrow the selection of machine learning algorithms to be tested. First, a specific literature on machine learning for cyberbullying detection is important in selecting a specified classifier. The pre-eminence of the classifier may be circumscribed to a given domain [134]. Therefore, general previous research and findings on machine learning can used as a guide to select a machine learning algorithm. Second, a literature review of text mining [135], [136] can be used as a guide. Third, a performance comparison of comprehensive datasets [137] can be used as basis to select machine learning algorithms. However, although these three points can be used as guide to narrow the selection of machine learning algorithms, researchers need to test many machine learning algorithms to identify the optimal classifier for an accurate predictive model.

E. IMBALANCED CLASS DISTRIBUTION

In many cases of real data, datasets naturally have imbalanced classes in which the normal class has a large number of instances and the abnormal class has a small number of instances in the dataset. Abnormal class instances are rare and difficult to be collected from real-world applications. Examples of imbalanced data applications are fraud detection, instruction detection, and medical diagnosis. Similarly, the number of cyberbullying posts is expected to be much less than the number of non-cyberbullying posts, and this assumption generates an imbalanced class distribution in the dataset in which the instances of non-cyberbullying contain much more posts than those of cyberbullying. Such cases can prevent the model from correctly classifying the examples. Many methods have been proposed to solve this issue, and examples include SMOTE [138] and weight adjustment (cost-sensitive technique) [139].

The SMOTE technique [138] is applied to avoid overfitting, which occurs when particular replicas of minority classes are added to the main dataset. A subdivision of data is reserved from the minority class as an example, and new synthetic similar classes are generated. These synthetic classes are then added to the original dataset. The created dataset is used to train the machine learning methods. The cost-sensitive technique is utilized to control the imbalance class [139]. It is based on creating is a cost matrix, which defines the costs experienced in false positives and false negatives.

F. EVALUATION METRIC SELECTION

Accuracy, precision, recall, and AUC are commonly used as evaluation metrics [19], [38]. Evaluation metric selection is important. The selection is based on the nature of manually labeled data. Selecting an inappropriate evaluation metric may result in better performance according to the selected evaluation metric. Then, the researcher may find the results to be significantly improved, although an investigation of how the machine learning model is evaluated may produce contradicting results and may not truly reflect the improvement of performance. For example, cyberbullying posts are commonly considered abnormal cases, whereas non-cyberbullying posts are considered normal cases. The ratio between cyberbullying and non-cyberbullying is normally large. Generally, non-cyberbullying posts comprise a large portion. For example, 1000 posts are manually labeled as cyberbullying and non-cyberbullying. The non-cyberbullying posts are 900, and the remaining 100 posts are cyberbullying. If a machine learning classifier classifies all 1000 posts as non-cyberbullying and is unable to classify any posts (0) as cyberbullying, then this classifier is considered impractical. By contrast, if researchers use accuracy as the main evaluation metric, then the accuracy of this classifier calculated as mentioned in the accuracy equation will yield a high accuracy percentage.

In the example, the classifier fails to classify any cyberbullying posts but obtains a high accuracy percentage. Knowing the nature of manually labeled data is important in selecting an evaluation metric. In cases where data are imbalanced, researchers may need to select AUC as the main evaluation metric. In class-imbalance situations, AUC is more robust than other performance metrics [140]. Cyberbullying and non-cyberbullying data are commonly imbalanced datasets (non-cyberbullying posts outnumber the cyberbullying ones) that closely represent the real-life data that machine learning algorithms need to train on. Accordingly, the learning performance of these algorithms is independent of data skewness [73]. Special care should be taken in selecting the main evaluation metric to avoid uncertain results and appropriately evaluate the performance of machine learning algorithms.

IV. ISSUES AND CHALLENGES

This section presents the issues and challenges while guiding future researchers to explore the domain of sentiment analysis through leveraging machine learning algorithms and models for detecting cyberbullying through social media.

A. HUMAN DATA CHARACTERISTICS

Although SM big data provide insights into large human behavior data, in reality, the analysis of such big data remains subjective [141]. Building human prediction systems involves steps where subjectivity about human behavior does exist. For example, when creating a manually labeled dataset to train a machine learning algorithm to predict cyberbullying posts, human bias may exist based on how cyberbullying is being defined and the criteria used to categorize the text as cyberbullying text.

Moreover, subjectivity may exist during the creation of a set of features (learning factors) in the feature engineering process. For example, the pre-processing stage involves a ‘‘data cleaning’’ process wherein choices about what features will be counted, and which will be ignored are constructed. This process is inherently subjective [141].

Predicting human behavior is crucial but complex. To achieve an effective prediction of human behavior, the patterns that exist and are used for constructing the prediction model should also exist in the future input data. The patterns should clearly represent features that occur in current and future data to retain the context of the model. Given that big data are not generic and dynamic in nature, the context of these data is difficult to understand in terms of scale and even more difficult to maintain when data are reduced to fit into a machine learning model. Handling context of big data is challenging and has been presented as an important future direction [141].

Furthermore, human behavior is dynamic. Knowing when online users change the way of committing cyberbullying is an important component in updating the prediction model with such changes. Therefore, dynamically updating the prediction model is necessary to meet human behavioral changes [1].

B. CULTURE EFFECT

What was considered cyberbullying yesterday might not be considered cyberbullying today, and what was previously considered cyberbullying may not be considered cyberbullying now due to the introduction of OSNs. OSNs have a globalized culture. However, machine learning always learns from the examples provided. Consequently, designing different examples that represent a different culture remains to be defined, and robust work from different disciplines is required. For this purpose, cross disciplinary coordination is highly desirable.

C. LANGUAGE DYNAMICS

Language is quickly changing, particularly among the young generation. New slang is regularly integrated into the language culture. Therefore, researchers are invited to propose dynamic algorithms to detect new slang and abbreviations related to cyberbullying behavior on SM websites and keep updating the training processes of machine learning algorithms by using newly introduced words.

D. PREDICTION OF CYBERBULLYING SEVERITY

The level of cyberbullying severity should be determined. The effect of cyberbullying is proportional to its severity and spread. Predicting different levels of cyberbullying severity does not only require machine learning understanding but also a comprehensive investigation to define and categorize the level of cyberbullying severity from social and psychological perceptions. Efforts from different disciplines are required to define and identify the levels of severity then introduce related factors that can be converted into features to build multi-classifier machine learning for classifying cyberbullying severity into different levels as opposed to a binary classifier that only detects whether an instance is cyberbullying or not.

E. UNSUPERVISED MACHINE LEARNING

Human learning is essentially unsupervised. The structure of the world was discovered by observing it and not by being told the name of every objective. Nevertheless, unsupervised machine learning has been overshadowed by the success of supervised learning [142]. This gap in literature may be caused by the fact that nearly all current studies rely on manually labeled data as the input to supervised algorithms for classifying classes. Thus, finding patterns between two classes by using unsupervised grouping remains difficult. Intensive research is required to develop unsupervised algorithms that can detect effective patterns from data. Traditional machine learning algorithms lack the capability to handle cyberbullying big data.

Deep learning has recently attracted the attention of many researchers in different fields. Natural language understanding is a new area in which deep learning is poised to make a large effect over the next few years [142].

The traditional machine learning algorithms pointed out in this survey lacks the capability to process big data in a standalone format. Big data have rendered traditional machine learning algorithms impotent. Cyberbullying big data generated from SM require advanced technology for the processing of the generated data to gain insights and help in making intelligent decisions.

Big data are generated at a very high velocity, variety, volume, verdict, value, veracity, complexity, etc. Researchers need to leverage various deep learning techniques for processing social media big data for cyberbullying behaviors. The deep learning techniques and architectures with a potential to explore the cyberbullying big data generated from SM can include generative adversarial network, deep belief network, convolutional neural network, stacked autoencoder, deep echo state network, and deep recurrent neural network. These deep learning architectures remain unexplored in cyberbullying detection in SM.

V. CONCLUSIONS AND FUTURE DIRECTIONS

This study reviewed existing literature to detect aggressive behavior on SM websites by using machine learning approaches. We specifically reviewed four aspects of detecting cyberbullying messages by using machine learning approaches, namely, data collection, feature engineering, construction of cyberbullying detection model, and evaluation of constructed cyberbullying detection models. Several types of discriminative features that were used to detect cyberbullying in online social networking sites were also summarized. In addition, the most effective supervised machine learning classifiers for classifying cyberbullying messages in online social networking sites were identified.

One of the main contributions of current paper is the definition of evaluation metrics to successfully identify the significant parameter so the various machine learning algorithms can be evaluated against each other. Most importantly we summarized and identified the important factors for detecting cyberbullying through machine learning techniques specially supervised learning. For this purpose, we have used accuracy, precision recall and f-measure which gives us the area under the curve function for modeling the behaviors in cyberbullying. Finally, the main issues and open research challenges were described and discussed.

Considerable research effort is required to construct highly effective and accurate cyberbullying detection models. We believe that the current study will provide crucial details on and new directions in the field of detecting aggressive human behavior, including cyberbullying detection in online social networking sites.

REFERENCES

[1] V. Subrahmanian and S. Kumar, "Predicting human behavior: The next frontiers," *Science*, vol. 355, no. 6324, p. 489, 2017.

[2] H. Lauw, J. C. Shafer, R. Agrawal, and A. Ntoulas, "Homophily in the digital world: A LiveJournal case study," *IEEE Internet Comput.*, vol. 14, no. 2, pp. 15–23, Mar./Apr. 2010.

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

[4] L. Phillips, C. Dowling, K. Shaffer, N. Hodas, and S. Volkova, "Using social media to predict the future: A systematic literature review," 2017, *arXiv:1706.06134*. [Online]. Available: https://arxiv.org/abs/1706.06134

[5] H. Quan, J. Wu, and Y. Shi, "Online social networks & social network services: A technical survey," in *Pervasive Communication Handbook*. Boca Raton, FL, USA: CRC Press, 2011, p. 4.

[6] J. K. Peterson and J. Densley, "Is social media a gang? Toward a selection, facilitation, or enhancement explanation of cyber violence," *Aggression Violent Behav.*, 2016.

[7] BBC. (2012). *Huge Rise in Social Media*. [Online]. Available: http://www.bbc.com/news/uk-20851797

[8] P. A. Watters and N. Phair, "Detecting illicit drugs on social media using automated social media intelligence analysis (ASMIA)," in *Cyberspace Safety and Security*. Berlin, Germany: Springer, 2012, pp. 66–76.

[9] M. Fire, R. Goldschmidt, and Y. Elovici, "Online social networks: Threats and solutions," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 2019–2036, 4th Quart., 2014.

[10] N. M. Shekokar and K. B. Kansara, "Security against sybil attack in social network," in *Proc. Int. Conf. Inf. Commun. Embedded Syst. (ICICES)*, 2016, pp. 1–5.

[11] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer, "Detecting and tracking political abuse in social media," in *Proc. 5th Int. AAAI Conf. Weblogs Social Media*, 2011, pp. 297–304.

[12] A. Aggarwal, A. Rajadesingan, and P. Kumaraguru, "PhishAri: Automatic realtime phishing detection on Twitter," in *Proc. eCrime Res. Summit (eCrime)*, Oct. 2012, pp. 1–12.

[13] S. Yardi *et al.*, "Detecting spam in a Twitter network," *First Monday*, Jan. 2009. [Online]. Available: https://firstmonday.org/article/view/2793/2431

[14] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu, "Analyzing spammers' social networks for fun and profit: A case study of cyber criminal ecosystem on twitter," in *Proc. 21st Int. Conf. World Wide Web*, 2012, pp. 71–80.

[15] G. R. S. Weir, F. Toolan, and D. Smeed, "The threats of social networking: Old wine in new bottles?" *Inf. Secur. Tech. Rep.*, vol. 16, no. 2, pp. 38–43, 2011.

[16] M. J. Magro, "A review of social media use in e-government," *Administ. Sci.*, vol. 2, no. 2, pp. 148–161, 2012.

[17] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improving cyberbullying detection with user context," in *Advances in Information Retrieval*. Berlin, Germany: Springer, 2013, pp. 693–696.

[18] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in *Proc. Int. Conf. Privacy, Secur., Risk Trust (PASSAT)*, Sep. 2012, pp. 71–80.

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

[20] W. Dong, S. S. Liao, Y. Xu, and X. Feng, "Leading effect of social media for financial fraud disclosure: A text mining based analytics," in *Proc. AMCIS*, San Diego, CA, USA, 2016.

[21] M. S. Rahman, T.-K. Huang, H. V. Madhyastha, and M. Faloutsos, "FRAppE: Detecting malicious Facebook applications," in *Proc. 8th Int. Conf. Emerg. Netw. Exp. Technol.*, 2012, pp. 313–324.

[22] S. Abu-Nimeh, T. Chen, and O. Alzubi, "Malicious and spam posts in online social networks," *Computer*, vol. 44, no. 9, pp. 23–28, Sep. 2011.

[23] B. Doerr, M. Fouz, and T. Friedrich, "Why rumors spread so quickly in social networks," *Commun. ACM*, vol. 55, no. 6, pp. 70–75, Jun. 2012.

[24] J. W. Patchin and S. Hinduja, *Words Wound: Delete Cyberbullying and Make Kindness Go Viral*. Golden Valley, MN, USA: Free Spirit Publishing, 2013.

[25] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Antisocial behavior in online discussion communities," in *Proc. 9th Int. AAAI Conf. Web Social Media*, Apr. 2015.

[26] S. Liu, J. Zhang, and Y. Xiang, "Statistical detection of online drifting Twitter spam: Invited paper," in *Proc. 11th ACM Asia Conf. Comput. Commun. Secur.*, 2016, pp. 1–10.

[27] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang, "Twitter spammer detection using data stream clustering," *Inf. Sci.*, vol. 260, pp. 64–73, Mar. 2014.

[28] M. Jiang, S. Kumar, V. S. Subrahmanian, and C. Faloutsos, "KDD 2017 tutorial: Data-driven approaches towards malicious behavior modeling," *Dimensions*, vol. 19, p. 42, 2017.

[29] S. Y. Jeong, Y. S. Koh, and G. Dobbie, "Phishing detection on Twitter streams," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*. Cham, Switzerland: Springer, 2016, pp. 141–153.

[30] I. Frommholz, H. M. Al-Khateeb, M. Potthast, Z. Ghasem, M. Shukla, and E. Short, "On textual analysis and machine learning for cyberstalking detection," *Datenbank-Spektrum*, vol. 16, no. 2, pp. 127–135, 2016.

[31] M. McCord and M. Chuah, "Spam detection on Twitter using traditional classifiers," in *Autonomic and Trusted Computing*. Berlin, Germany: Springer, 2011, pp. 175–186.

[32] X. Chen, R. Chandramouli, and K. P. Subbalakshmi, "Scam detection in Twitter," in *Data Mining for Service*. Berlin, Germany: Springer, 2014, pp. 133–150.

[33] A. H. Wang, "Detecting spam bots in online social networking sites: A machine learning approach," in *Data and Applications Security and Privacy XXIV*. Berlin, Germany: Springer, 2010, pp. 335–342.

[34] X. Zheng, Z. Zeng, Z. Chen, Y. Yu, and C. Rong, "Detecting spammers on social networks," *Neurocomputing*, vol. 159, pp. 27–34, Jul. 2015.

[35] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 3, p. 18, 2012.

[36] M. Dadvar and F. De Jong, "Cyberbullying detection: A step toward a safer Internet yard," in *Proc. 21st Int. Conf. Companion World Wide Web*, 2012, pp. 121–126.

[37] S. O. Sood, J. Antin, and E. Churchill, "Using crowdsourcing to improve profanity detection," in *Proc. AAAI Spring Symp., Wisdom Crowd*, 2012, pp. 69–74.

[38] P. Galán-García, J. G. de la Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in Twitter social network: Application to a real case of cyberbullying," in *Proc. Int. Joint Conf. SOCO-CISIS-ICEUTE*. Cham, Switzerland: Springer, 2014, pp. 419–428.

[39] Q. Huang, V. K. Singh, and P. K. Atrey, "Cyber bullying detection using social and textual analysis,"in *Proc. 3rd Int. Workshop Socially-Aware Multimedia*, 2014, pp. 3–6.

[40] R. M. Kowalski, *Cyberbullying: Bullying in the Digital Age*. Hoboken, NJ, USA: Wiley, 2012.

[41] T. Nakano, T. Suda, Y. Okaie, and M. J. Moore, "Analysis of cyber aggression and cyber-bullying in social networking," in *Proc. IEEE 10th Int. Conf. Semantic Comput. (ICSC)*, Feb. 2016, pp. 337–341.

[42] G. S. O'Keeffe and K. Clarke-Pearson, "The impact of social media on children, adolescents, and families," *Pediatrics*, vol. 127, no. 4, pp. 800–804, 2011.

[43] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2012, pp. 656–666.

[44] R. M. Kowalski and S. P. Limber, "Psychological, physical, and academic correlates of cyberbullying and traditional bullying," *J. Adolescent Health*, vol. 53, no. 1, pp. S13–S20, 2013.

[45] H. Sampasa-Kanyinga, P. Roumeliotis, and H. Xu, "Associations between cyberbullying and school bullying victimization and suicidal ideation, plans and attempts among Canadian schoolchildren," *PLoS ONE*, vol. 9, no. 7, 2014, Art. no. e102145.

[46] S. Hinduja and J. W. Patchin, "Bullying, cyberbullying, and suicide," *Arch. Suicide Res.*, vol. 14, no. 3, pp. 206–221, 2010.

[47] R. Slonje, P. K. Smith, and A. Frisén, "The nature of cyberbullying, and strategies for prevention," *Comput. Hum. Behav.*, vol. 29, no. 1, pp. 26–32, 2013.

[48] E. Whittaker and R. M. Kowalski, "Cyberbullying via social media," *J. School Violence*, vol. 14, no. 1, pp. 11–29, 2015.

[49] F. Sticca and S. Perren, "Is cyberbullying worse than traditional bullying? Examining the differential roles of medium, publicity, and anonymity for the perceived severity of bullying," *J. Youth Adolescence*, vol. 42, no. 5, pp. 739–750, 2013.

[50] S. Wen, J. Jiang, Y. Xiang, S. Yu, W. Zhou, and W. Jia, "To shut them up or to clarify: Restraining the spread of rumors in online social networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 12, pp. 3306–3316, Dec. 2014.

[51] K. Van Royen, K. Poels, W. Daelemans, and H. Vandebosch, "Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability," *Telematics Inform.*, vol. 32, no. 1, pp. 89–97, 2015.

[52] K. Van Royen, K. Poels, and H. Vandebosch, "Harmonizing freedom and protection: Adolescents' voices on automatic monitoring of social networking sites," *Children Youth Services Rev.*, vol. 64, pp. 35–41, May 2016.

[53] R. M. Kowalski, "Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth," *Psychol. Bull.*, vol. 140, no. 4, pp. 1073–1137, 2014.

[54] Q. Li, "New bottle but old wine: A research of cyberbullying in schools," *Comput. Hum. Behav.*, vol. 23, no. 4, pp. 1777–1791, 2007.

[55] R. S. Tokunaga, "Following you home from school: A critical review and synthesis of research on cyberbullying victimization," *Comput. Hum. Behav.*, vol. 26, no. 3, pp. 277–287, May 2010.

[56] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97–107, Jan. 2014.

[57] Y. Liu, J. Yang, Y. Huang, L. Xu, S. Li, and M. Qi, "MapReduce based parallel neural networks in enabling large scale machine learning," *Comput. Intell. Neurosci.*, vol. 2015, p. 1, Jan. 2015.

[58] C. Wu, R. Buyya, and K. Ramamohanarao, "Big data analytics = machine learning + cloud computing," 2016, *arXiv:1601.03115*. [Online]. Available: https://arxiv.org/abs/1601.03115

[59] Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "A survey on deep learning for big data," *Inf. Fusion*, vol. 42, pp. 146–157, Jul. 2018.

[60] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 137–144, 2015.

[61] M. Dadvar, F. D. Jong, R. Ordelman, and D. Trieschnigg, "Improved cyberbullying detection using gender information," in *Proc. 25th Dutch-Belgian Inf. Retr. Workshop*, 2012, pp. 1–3.

[62] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Detection of cyberbullying incidents on the instagram social network," 2015, *arXiv:1503.03909*. [Online]. Available: https://arxiv.org/abs/1503.03909

[63] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, pp. 1289–1305, Mar. 2003.

[64] P. D. Turney, "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews," in *Proc. ACL*, 2002, pp. 417–424.

[65] R. M. Tong, "An operational system for detecting and tracking opinions in on-line discussion," in *Proc. Notes ACM SIGIR Workshop Oper. Text Classification*, 2001.

[66] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in *Proc. 8th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 1997, pp. 174–181.

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

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

[69] H. Chen, S. Mckeever, and S. J. Delany, "Harnessing the power of text mining for the detection of abusive content in social media," in *Advances in Computational Intelligence Systems*. Cham, Switzerland: Springer, 2017, pp. 187–205.

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

[71] V. Nahar, X. Li, and C. Pang, "An effective approach for cyberbullying detection," *Commun. Inf. Sci. Manage. Eng.*, vol. 3, no. 5, p. 238, 2013.

[72] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, and V. Hoste, "Detection and fine-grained classification of cyberbullying events," in *Proc. Int. Conf. Recent Adv. Natural Lang. Process. (RANLP)*, 2015, pp. 672–680.

[73] A. Mangaonkar, A. Hayrapetian, and R. Raje, "Collaborative detection of cyberbullying behavior in Twitter data," in *Proc. IEEE Int. Conf. Electro/Inf. Technol. (EIT)*, Dekalb, IL, USA, May 2015, pp. 611–616.

[74] H. Sanchez and S. Kumar, "Twitter bullying detection," Tech. Rep. UCSC ISM245, 2011.

[75] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan, "An introduction to MCMC for machine learning," *Mach. Learn.*, vol. 50, nos. 1–2, pp. 5–43, 2003.

[76] M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," *Nature Rev. Genet.*, vol. 16, no. 6, pp. 321–332, 2015.

[77] P. Langley and H. A. Simon, "Applications of machine learning and rule induction," *Commun. ACM*, vol. 38, no. 11, pp. 54–64, 1995.

[78] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," *Nature*, vol. 521, no. 7553, pp. 452–459, May 2015.

[79] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[80] Y. Liu, C. Kliman-Silver, and A. Mislove, "The tweets they are a-changin': Evolution of twitter users and behavior," in *Proc. Int. AAAI Conf. Weblogs Social Media (ICWSM)*, 2014, pp. 305–314.

[81] S. González-Bailón, N. Wang, A. Rivero, J. Borge-Holthoefer, and Y. Moreno, "Assessing the bias in samples of large online networks," *Social Netw.*, vol. 38, pp. 16–27, Jul. 2014.

[82] T. Cheng and T. Wicks, "Event detection using Twitter: A spatio-temporal approach," *PLoS ONE*, vol. 9, no. 6, p. e97807, 2014.

[83] A. Bellmore, A. J. Calvin, J.-M. Xu, and X. Zhu, "The five W's of 'bullying' on Twitter: Who, what, why, where, and when," *Comput. Hum. Behav.*, vol. 44, pp. 305–314, Mar. 2015.

[84] H. Margono, X. Yi, and G. K. Raikundalia, "Mining Indonesian cyber bullying patterns in social networks," in *Proc. 37th Australas. Comput. Sci. Conf.*, vol. 147, Australian Computer Society, 2014, pp. 115–124.

[85] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," in *Proc. 17th Int. Conf. Distrib. Comput. Netw.*, 2016, Art. no. 43.

[86] Á. García-Recuero, "Discouraging abusive behavior in privacy-preserving online social networking applications," in *Proc. 25th Int. Conf. Companion World Wide Web*, International World Wide Web Conferences Steering Committee, 2016, pp. 305–309.

[87] Y. Anzai, *Pattern Recognition and Machine Learning*. Amsterdam, The Netherlands: Elsevier, 2012.

[88] E. Calvete, I. Orue, A. Estévez, L. Villardón, and P. Padilla, "Cyberbullying in adolescents: Modalities and aggressors' profile," *Comput. Hum. Behav.*, vol. 26, no. 5, pp. 1128–1135, 2010.

[89] H. Vandebosch and K. Van Cleemput, "Cyberbullying among youngsters: Profiles of bullies and victims," *New Media Soc.*, vol. 11, no. 8, pp. 1349–1371, 2009.

[90] R. Slonje and P. K. Smith, "Cyberbullying: Another main type of bullying?" *Scand. J. Psychol.*, vol. 49, no. 2, pp. 147–154, 2008.

[91] K. R. Williams and N. G. Guerra, "Prevalence and predictors of Internet bullying," *J. Adolescent Health*, vol. 41, no. 6, pp. S14–S21, 2007.

[92] O. T. Arıcak, "Psychiatric symptomatology as a predictor of cyberbullying among University Students," *Eurasian J. Educ. Res.*, vol. 34, no. 1, p. 169, 2009.

[93] I. Connolly and M. O'Moore, "Personality and family relations of children who bully," *Personality Individual Differences*, vol. 35, no. 3, pp. 559–567, 2003.

[94] L. Corcoran, I. Connolly, and M. O'Moore, "Cyberbullying in Irish schools: An investigation of personality and self-concept," *Irish J. Psychol.*, vol. 33, no. 4, pp. 153–165, 2012.

[95] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in *Proc. 5th Int. AAAI Conf. Weblogs Social Media*, 2011, pp. 11–17.

[96] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of harassment on Web 2.0," in *Proc. Content Anal. Web*, 2009, pp. 1–7.

[97] M. Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka, and K. Araki, "Machine learning and affect analysis against cyber-bullying," in *Proc. 36th AISB*, 2010, pp. 7–16.

[98] E. Raisi and B. Huang, "Cyberbullying identification using participant-vocabulary consistency," 2016, *arXiv:1606.08084*. [Online]. Available: https://arxiv.org/abs/1606.08084

[99] A. Squicciarini, S. Rajtmajer, Y. Liu, and C. Griffin, "Identification and characterization of cyberbullying dynamics in an online social network," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, 2015, pp. 280–285.

[100] R. M. Gray, "Entropy and information," in *Entropy and Information Theory*. New York, NY, USA: Springer, 1990, pp. 21–55.

[101] I. Qabajeh and F. Thabtah, "An experimental study for assessing email classification attributes using feature selection methods," in *Proc. 3rd Int. Conf. Adv. Comput. Sci. Appl. Technol. (ACSAT)*, Dec. 2014, pp. 125–132.

[102] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*. Berlin, Germany: Springer, 2009, pp. 1–4.

[103] M. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proc. 17th Int. Conf. Mach. Learn.*, 2000, pp. 359–366.

[104] Z. Zheng, X. Wu, and R. Srihari, "Feature selection for text categorization on imbalanced data," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 80–89, 2004.

[105] D. H. Wolper and W. G. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, Apr. 1997.

[106] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1153–1176, 2nd Quart., 2016.

[107] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. Eur. Conf. Mach. Learn.* Berlin, Germany: Springer, 1998.

[108] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Dept. Comput. Sci., Nat. Taiwan Univ., Tech. Rep., 2003. [Online]. Available: http://www.csie. ntu.edu.tw/*cjlin/papers/guide/guide.pdf

[109] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, pp. 45–66, Nov. 2001.

[110] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 2013.

[111] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. Workshop Comput. Learn. Theory*, 1992, pp. 144–152.

[112] A. McCallum and K. Nigam, "A comparison of event models for naive Bayes text classification," in *Proc. AAAI Workshop Learn. Text Categorization*, 1998, pp. 1–8.

[113] H. Zhang, "The optimality of naive Bayes," Tech. Rep., 2004.

[114] H. Zhang, "The optimality of naive Bayes," in *Proc. IAAA*, vol. 1, no. 2, 2004, p. 3.

[115] N. Bora, V. Zaytsev, Y.-H. Chang, and R. Maheswaran "Gang networks, neighborhoods and holidays: Spatiotemporal patterns in social media," in *Proc. Int. Conf. Social Comput. (SocialCom)*, Sep. 2013, pp. 93–101.

[116] A. H. Wang, "Don't follow me: Spam detection in Twitter," in *Proc. Int. Conf. Secur. Cryptogr. (SECRYPT)*, Jul. 2010, pp. 1–10.

[117] D. M. Freeman, "Using naive bayes to detect spammy names in social networks," in *Proc. ACM Workshop Artif. Intell. Secur.*, 2013, pp. 3–12.

[118] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[119] D. R. Cutler, D. R. Cutler, T. C. Edwards, Jr., K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler, "Random forests for classification in ecology," *Ecology*, vol. 88, no. 11, pp. 2783–2792, 2007.

[120] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, 2008.

[121] P. Soucy and G. W. Mineau, "A simple KNN algorithm for text categorization," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov./Dec. 2001, pp. 647–648.

[122] Z. Deng, X. Zhu, D. Cheng, M. Zong, and S. Zhang, "Efficient kNN classification algorithm for big data," *Neurocomputing*, vol. 195, pp. 143–148, Jun. 2016.

[123] S. Dreiseitl, L. Ohno-Machado, H. Kittler, S. Vinterbo, H. Billhardt, and M. Binder, "A comparison of machine learning methods for the diagnosis of pigmented skin lesions," *J. Biomed. Informat.*, vol. 34, no. 1, pp. 28–36, 2001.

[124] D. W. Hosmer, Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, vol. 398. Hoboken, NJ, USA: Wiley, 2013.

[125] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. IJCAI*, 1995, pp. 1137–1145.

[126] P. K. Smith, *The Nature of School Bullying: A Cross-National Perspective*. London, U.K.: Psychology Press, 1999.

[127] J. J. Dooley, J. Pyżalski, and D. Cross, "Cyberbullying versus face-to-face bullying: A theoretical and conceptual review," *Zeitschrift Psychologie/J. Psychol.*, vol. 217, no. 4, pp. 182–188, 2009.

[128] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, "Is the sample good enough? Comparing data from twitter's streaming API with Twitter's firehose," 2013, *arXiv:1306.5204*. [Online]. Available: https://arxiv.org/abs/1306.5204

[129] *From IHML (I Hate My Life) to MOS (Mum Over Shoulder): Why This Guide to Cyber-Bullying Slang May Save Your Child's Life*, Dailymail, USA, 2014. [Online]. Available: https://www.dailymail.co.uk/news/article-2673678/Why-guide-cyber-bullying-slang-save-childs-life-From-IHML-I-hate-life-Mos-mum-shoulder.html

[130] J. N. Navarro and J. L. Jasinski, "Going Cyber: Using routine activities theory to predict cyberbullying experiences," *Sociol. Spectr.*, vol. 32, no. 1, pp. 81–94, 2012.

[131] C. Peersman, W. Daelemans, and L. Van Vaerenbergh, "Predicting age and gender in online social networks," in *Proc. 3rd Int. Workshop Search Mining User-Generated Contents*, 2011, pp. 37–44.

[132] D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder, "'How old do you think I Am?' A study of language and age in Twitter," in *Proc. 7th Int. AAAI Conf. Weblogs Social Media*, AAAI Press, 2013, pp. 439–448.

[133] V. Balakrishnan, "Cyberbullying among young adults in Malaysia: The roles of gender, age and Internet frequency," *Comput. Hum. Behav.*, vol. 46, pp. 149–157, May 2015.

[134] N. Maciá, E. Bernadó-Mansilla, A. Orriols-Puig, and T. K. Ho, "Learner excellence biased by data set selection: A case for data characterisation and artificial data sets," *Pattern Recognit.*, vol. 46, no. 3, pp. 1054–1066, Mar. 2013.

[135] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[136] V. Korde and C. N. Mahender, "Text classification and classifiers: A survey," *Int. J. Artif. Intell. Appl.*, vol. 3, no. 2, p. 85, 2012.

[137] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *J. Mach. Learn. Res*, vol. 15, no. 1, pp. 3133–3181, 2014.

[138] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

[139] X.-Y. Liu and Z.-H. Zhou, "The influence of class imbalance on cost-sensitive learning: An empirical study," in *Proc. 6th Int. Conf. Data Mining (ICDM)*, Dec. 2006, pp. 970–974.

[140] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[141] D. Boyd and K. Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," *Inf., Commun. Soc.*, vol. 15, no. 5, pp. 662–679, 2012.

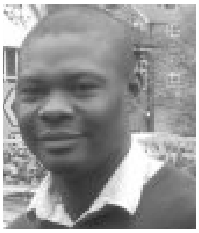
[142] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

MOHAMMED ALI AL-GARADI received the Ph.D. degree from the Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. He has published several articles in academic journals indexed in well-reputed databases such as ISI and Scopus. His research interests include cybersecurity, online social networking, machine learning text mining, deep learning, and the IoT.

MOHAMMAD RASHID HUSSAIN received the Ph.D. degree in information technology from Babasaheb Bhimrao Ambedkar Bihar University, India. He is currently an Assistant Professor with the College of Computer Science, King Khalid University, Abha, Saudi Arabia. His research interests include educational development & review, educational data mining, cloud intelligence, and mobile cloud computing and optimizations.

NAWSHER KHAN received the Ph.D. degree from University Malaysia Pahang (UMP), Malaysia, in 2013. He was a Postdoctoral Research Fellow with the University of Malaya (UM), Malaysia, in 2014. In 2005, he was appointed in National Database and Registration Authority (NADRA) under the Interior Ministry of Pakistan. In 2008, he has worked in National Highways Authority (NHA). He has served at Abdul Wali Khan University Mardan, Pakistan, as an Assistant Professor for 3 years, from 2014 to 2017. He is currently serving as an Associate Professor and the Director of Research Center, College of Computer Science, King Khalid University, Abha, Saudi Arabia. He has published more than 50 articles in various International Journals and Conference proceedings. His research interests include big data, cloud computing, data management, distributed systems, scheduling, replication, and sensor networks.

GHULAM MURTAZA is currently pursuing the Ph.D. degree with the Faculty of Computer Science and Information Technology, University of Malaya, Malaysia. He is also an Assistant Professor with Sukkur IBA University, Sukkur, Pakistan. He is currently on study leave to pursue his Ph.D. He has published several articles in well-reputed databases. His research interests include machine learning, deep learning, digital image processing, big data, and information retrieval.



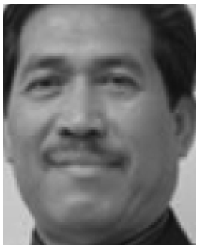
HENRY FRIDAY NWEKE received the B.Sc. degree in computer science from Ebonyi State University, Nigeria, and the M.Sc. degree in computer science from the University of Bedfordshire, U.K. He is currently pursuing the Ph.D. degree with the Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. His research interests include machine learning, deep learning, biomedical sensor analytics, human activity recognition, multi-sensor fusion, cloud computing, wireless sensor technologies, and emerging technology.

IHSAN ALI received the M.Sc. degree from Hazara University Manshera, Pakistan, in 2005, and the M.S. degree in computer system engineering from the GIK Institute, in 2008. He is currently pursuing the Ph.D. degree with the Faculty of Computer Science and Information Technology, University of Malaya. He has more than five years teaching and research experience in different country, including Saudi Arabia, USA, Pakistan, and Malaysia. He has served as a Technical Program Committee Member for the IWCMC 2017, AINIS 2017, Future 5V 2017, and also an Organizer of Special session on fog computing in Future 5V 2017. He has published more than 30 papers in the international journals and conferences. His research interests include wireless sensor networks, underwater sensor networks, sensor cloud, fog computing, and the IoT. He is also an Active Member of the IEEE, ACM, the International Association of Engineers (IAENG), and the Institute of Research Engineers and Doctors (the IRED). He is also a Reviewer of *Computers & Electrical Engineering*, *KSII Transactions on Internet and Information Systems*, *Mobile Networks and Applications*, the *International Journal of Distributed Sensor Networks*, the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, *Computer Networks*, IEEE ACCESS, FGCS, and the *IEEE Communication Magazine*.

GHULAM MUJTABA received the master's degree in computer science from FAST National University, Karachi, Pakistan, and the Ph.D. degree from the Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. He has received the gold medal for the masterâĂŹs degree. He has been an Assistant Professor with Sukkur IBA University, Sukkur, Pakistan, since 2006. He has vast experience in teaching and research. Before he joined Sukkur IBA University, he was with a well-known software house in Karachi for four years. He has also published several articles in academic journals indexed in well-reputed databases such as ISI and Scopus. His research interests include machine learning, online social networking, text mining, deep learning, and information retrieval.

HARUNA CHIROMA received the B.Tech. degree from Abubakar Tafawa Balewa University, the M.Sc. degree from Bayero University Kano, and the Ph.D. degree from the University of Malaya, all in computer science, and the Post-Graduate Diploma in education from Usman Danfodio University. He was a Visiting Senior Lecturer with Abubakar Tafawa Balewa University, Bauchi, and a Lecturer with the Federal College of Education, Gombe. As a teacher, he has developed interest in advanced learning technology. He has published more than 100 academic articles in international refereed ISI WoS Journals, Edited Books, Conference Proceedings, and Local Journals. He participated in the 2017 and 2018 QS world universities ranking evaluation of world universities research strengths in computer science. His research interests include deep learning, nature-inspired algorithms for machine learning, with special focus on their applications to internet of vehicles, autonomous vehicles, the Internet of Things, big data analytics, edge computing, cybersecurity, fog computing, and cloud computing. He has served in various capacities in more than 20 international conferences across the world. He is a member of the ACM, INNS, NCS, IAENG, and the Association for Computing Machinery (ACM). He is an Associate Editor of the IEEE ACCESS and ISI WoS indexed Journal. He is a Leading Volume Editor of an edited book *Advances on Computational Intelligence in Energy-the Applications of Nature-Inspired & Metaheuristic Algorithms in Energy* (Heidelberg, Berlin: Springer), renowned series of Lecture Notes in Energy. He is a Reviewer for 15 ISI WoS indexed journals, such as *Applied Energy Q1* (Elsevier), *Applied Soft Computing Q1* (Elsevier), *Knowledge Based System Q1* (Elsevier), *Energy and Building Q1* (Elsevier), *Neural Computing and Applications Q1* (Springer), the *Journal of the Operational Research Society* (Springer), and *PLOS One*.

HASAN ALI KHATTAK received the B.Sc. degree in computer science from the University of Peshawar, Peshawar, Pakistan, in 2006, the master's degree in information engineering from the Politecnico di Torino, Torino, Italy, in 2011, and the Ph.D. degree in electrical and computer engineering from the Politecnico di Bari, Bari, Italy, in 2015. He has been serving as an Assistant Professor of computer science, since 2016. He is involved in a number of funded research projects in the Internet of Things, semantic web, and fog computing while exploring Ontologies, Web Technologies using Contiki OS, NS 2/3, and Omnet++ frameworks. His current research interests include distributed systems, web of things and vehicular ad hoc networks, and data and social engineering for smart cities.



ABDULLAH GANI received the Diploma degree in computer science from ITM, the B.Phil. and M.Sc. degrees in information management from the University of Hull, U.K., and the Ph.D. degree in computer science from the University of Sheffield, U.K. He acquired the Teaching Certificate from Kinta Teaching College, Ipoh. He has vast teaching experience due to having worked in a number of educational institutions locally and abroad–schools, the Malay Women Teaching College, Melaka, Ministry of Education, the Rotterham College of Technology and Art, Rotterham, U.K., and the University of Sheffield. He is currently a Professor with the Department of Computer System and Technology, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia. Since then, more than 150 academic papers have been published in proceedings and respectable journals internationally within top 10% ranking. He received a very good number of citation in Web of Science and Scopus databases and actively supervises. His interest in research kicked off, in 1983 when he was chosen to attend the 3 month Scientific Research Course in RECSAM by the Ministry of Education, Malaysia.