ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940www.ijasem.org

Vol 9, Issuse.1 March 2021

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

G.Harika[1],B.Venkateswarlu[2],Dr.M.Bal Raju[3],P.Karunakar Reddy[4],

ABSTRACT

It wasn't until the advent of information and communication technology (ICT) that social contacts began to expand beyond geographical bounds. With the recent advancements in communication technology, the time and space constraints of conventional communication are no longer an issue. They have ushered in a new age of user-generated content, online networks of people, and statistics on human activity. Social media (SM) platforms, in particular, have been misused to develop a new type of anger and violence that happens only online. This study examines a brand-new method for displaying hostile conduct on social media websites. The development of predictive algorithms to combat aggressive behaviour in SM is also explained in this section. Cyberbullying prediction models in SM have a number of challenges, which we address in our evaluation of cyberbullying prediction models. When it comes to cyberbullying detection, this document gives an overview of the whole procedure. When it comes to predicting cyberbullying behaviours, several machine learning methods are being used, however the main focus is on feature selection techniques and their subsequent application to data collecting and feature engineering. New study avenues have also been identified as a result of the highlighted concerns and limitations.

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

INTRODUCTION

To better predict and detect the detrimental consequences of big data, researchers with limited resources may employ machine or deep learning methods [1]. People and human behaviour, especially cyberbullying [3] are covered extensively. Although it's now feasible to do a massive data analysis, it may also reveal previously unreachable knowledge using deep learning from this acquisition. If [1] is the case, thenBig data analytics has improved the quality of social media and other human-related data sources (SM).Even the capability of foreseeing the future has become a fact of everyday life. We may use machine learning algorithms to analyse SM data and integrate it with huge data in order to predict the future of different algorithms. A method for detecting and preventing hostile behaviour must be devised by analysing data on human behaviour and interactivity.To ensure this article was approved for publication, Kathiravan Srinivasan, the article's associate editor, was in charge of supervising its examination and approval. theorem fusion, many faces and angles, and as well as methodologies drawn from a wide range of academic fields. The availability of large-scale information generates new research problems, novel computational tools, transdisciplinary approaches, and excellent opportunities to statistically investigate many vital concerns. For example, traditional statistical techniques are difficult to scale and are inaccurate in this context. Structured data on human behaviour and small-scale human networks are often employed to assist these tactics in practise (traditional social networks). As a consequence, implementing these strategies on massive online social networks is fraught with difficulties (OSN). Due to OSNs' fast growth, they both encourage and facilitate the spread of violent behaviour.

Professor[1,2,3,4,] Assistant Professor[1,2,3,4], ,Associate Professor[1,2,3,4],
Department of CSE Engineering,
Pallavi Engineering College,
Mail ID:harikagora@gmail.com, Mail ID:bvenkat1109@gmail.com,
Kuntloor(V),Hayathnagar(M),Hyderabad,R.R.Dist.-501505.

As for OSNs, researchers may use the information they give in the form of OSN data to develop effective techniques to recognise and restrict improper conduct and/or aggressive behaviour. It is possible for criminals to carry out hostile activities and to create networks for the purpose of doing crime via the usage of OSN. Complex systems demand tactics that include both the content and the network when trying to detect and restrict aggressive behaviour.Because of this, the remainder of this paper is organised as follows. Users may participate in abusive behaviour in a new manner on social media networks, as detailed in Subsection I.A. Reasons for constructing prediction models to counteract SM hostility are laid forth in this section. To I.C., building cyberbullying prediction models is of paramount significance. This study's methodology is explained in detail by I.D, so please elaborate. Section 2 examines in depth the cyberbullying prediction models of SM websites, from data collection to evaluation. Section 3 provides a wealth of information on how to build a cyberbullying prediction model. Research concerns and prospective future research projects are discussed in the paper's fourth part.The use of small tools has become more combative.The way we interact with one another has been transformed by advances in communication technologies. Advances in communication technology have made it possible for people to communicate with one other at any time and from any location. In the last several years, social media websites, blogs, online forums, and online sharing platforms have grown in popularity. In the digital era, aggression and violence take on new forms[6]. There has been a huge surge in aggressive activity on social media (SM) during the last several years. Social media websites (SM), which are often accessed through mobile devices, have had a profound impact on user experience. Since social media sites like Facebook and Twitter are free and open to everyone, anybody may post and write on them, no matter where they are located. Users may easily engage in aggressive activity as a result. SM websites are used by hundreds of thousands of individuals every day throughout the world. Ideas, views, tastes, attitudes, and dialogues may be quickly shared through social media. Users' interactions with one other on social media may help us learn about human behaviour patterns [11]. It is now easier than ever to study social interaction trends due of SM websites.Others may engage in a particularly extreme kind of unethical behaviour because to SM websites, which modernise the instruments for persuading people. Online complex networks such as this have seen a major transformation in the last decade due to the rise of social media communication. You no longer only communicate with individuals you know and those you've never met; online communication has developed into a source of pleasure in and of itself. Despite the various benefits that SM websites provide its users, cyber criminals may utilise them for a range of illegal or unfriendly activities. On OSN sites, aggressive activity and/or misbehaviour such as cyberbullying, phishing, spam distribution, malware distribution, and cyberbullying are common. It's not uncommon to observe people engaging in aggressive behaviour on social media (SM). [9] and [15] are two of the most influential social media platforms for encouraging violent behaviour..In the Web 2.0 world, OSN communication is a hot new trend. Users are more likely to retain an online presence as a consequence of the added features of Web 2.0, thanks to the development of profiles and pages. Internet 1.0 restricted users to passive consumption of content, whereas Web 2.0 makes it possible for people to actively create and share their own. The trademarks of SM websites are coordination, participation, empowerment, and timeliness [16]. Criminals may use social media networks as a forum to perpetrate harmful behaviour without directly addressing the victims. All of these types of hostile behaviour, including cyberbullying and financial fraud, as well as destructive software and social engineering and phishing, are forms of the Internet.The second [iii].] Information may be transferred and shared on a social media site (SM site). All of these applications allow you to exchange messages, links, photos, and videos. Since social media links billions of individuals globally, it has become a medium for a broad spectrum of violent and antagonistic activity. A substantial number of individuals are contacted by cybercriminals via social media platforms.

USER ACTIONS ON A WEBSITE There has been a lot of interest in machine learning algorithms in recent years. Machine learning research has produced a plethora of models, tools, and algorithms for dealing with massive amounts of data in order to solve real-world problems [24, 25]. Machine learning algorithms have been used extensively to analyse spam detection, phishing detection, and cyberbullying detection [26–28] and [29–30]. Aggressive behaviour may take many forms, including spam, phishing, the spread of viruses, and cyberbullying. Textual cyberbullying has been the most prevalent kind of hostile behaviour on SM websites due to the flexibility that users have to post on their platforms [17], [35]–[39].On SM websites, there is more than just text and/or non-text content concerning aggressive action. This study's predictions concerning aggressive behaviour are based on an examination of social media material. Textual OSN content is all that can be used in this kind of analysis to predict cyberbullying behaviour. Because of its simplicity of use, cyberbullying is both dangerous and fast spreading. Using a laptop or mobile phone with an Internet connection, bullies don't need to approach their victims to conduct misdeeds. SM websites have become more popular, which has led to an increase in online bullying. Because of the way they are set up, social media platforms are ideal breeding grounds for cyberbullying. For cyberbullying, email and text messages may be used, but only on a limited number of people. Since a result of this, cyberbullying may spread across geographical boundaries, as users can form connections and interact with one other regardless of where they are situated. SMwebsites' increasingly antagonistic user behaviour has been traced back to this [41]. Thus, the creation of an effective model for forecasting cyberbullying is of practical significance. All of these indicators are taken into consideration in this study's content-based approach to predicting textual cyberbullying on SM websites. The reason for doing this review will be discussed in the next section.

IT IS ESSENTIAL TO CREATE MODELS FOR PREDICTING CYBERBULLYING.

A number of factors prompted the researchers to conduct this study on social media and cyberbullying. While there's no denying the prevalence of cyberbullying, it's also been recognised as a serious public health problem [43]. Psychological and physical health problems, as well as academic performance, have been linked to cyberbullying in studies [44]. Victims of cyberbullying are more likely to consider suicide, according to research [45, 46]. Other research [45, 46] identified a correlation between cyberbullying and suicidal ideation. Identifying aggressive behaviour linked to human security is more important than a prediction model for aggressive behaviour linked to machine security in the context of cyberbullying.Cyberbullying may occur at any time and in any location. It's difficult to avoid becoming a victim of cyberbullying since it may happen at any time and from any place. This kind of crime might be committed via the use of public commenting and status updates. Victims have little power to halt these atrocities. [47] In spite of the fact that social media sites have become an integral part of many people's daily lives, cyberbullying victimisation is most common on these platforms [48]. In order to reach a large audience while protecting user privacy, social media (SM) platforms like Twitter are widely recognised [9]. Public cyberbullying is more harmful than private, and anonymous cyberbullying is more harmful than non-anonymous [49], [50]. So, cyberbullying on social media platforms, which allow for public and anonymous cyberbullying scenarios, has gotten more serious. Because of these characteristics, social media networks like Twitter pose a significant threat to victims of cyberbullying [43].Cyberbullying should be routinely tracked, according to recent research [51]. Using traditional approaches to combat cyberbullying is no longer successful in the age of big data and social networks, according to a study of 14 groups of high school students. Automated monitoring based on machine learning is also required to deal with large amounts of complex data.Now, 2.5 quintillion bytes per day [56] is produced. Every day, businesses generate enormous volumes of data. From the Internet, social networks, and sensors, huge datasets are generated [57]. Volume, diversity, variability and complexity are only a few of the nine characteristics of big data that are often known as the "big nine" [58]. Each day, Flickr generates roughly 3.6TB of data; this is more than double the amount of data that Google handles; and the Internet collects an estimated 1.8PB of data every single day.Members of SM have access to a wide range of tools, including the ability to share and exchange information and files. Businesses, people, and items all exchange information on social media platforms, resulting in a massive amount of data. There are a variety of social media platforms available, two examples of which are Facebook and YouTube. SM outlets may generate both organised and unstructured data. SM analytics focuses on the structured and unstructured data generated by SM outlets. Link prediction, community, content, social effect, and social influence analysis may all be done using both structured and unstructured techniques to SM analytics. SM has entered the big data era. For example, Facebook has over 20 PB of storage capacity for its 260 billion images and can process up to one million photos each second. Facebook. A total of 60 hours of video are downloaded from YouTube per minute [60].Developing machine learning classifiers from tagged text occurrences [19], [38], [61]–[63] is the most common method for predicting cyberbullying. The semantic orientation of words and phrases in a document is also taken into consideration by lexicon-based models when computing text orientation. [64] In lexicon-based models, lexicons may be constructed manually (as in [65]) or automatically (as in [66]). The lexicon-based approach to cyberbullying prediction is unusual in the literature, however. For the most part, this is due to the fact that the texts on SM websites are written in an unstructured form, making it hard for lexicon-based techniques to detect cyberbullying based on lexicons [67–69]. Machine learning algorithms, on the other hand, often utilise lexicons to extract characteristics, which are often employed as inputs. The number of obscene phrases in a communication may be identified by employing a profanity-based lexicon, for example, as a profanity feature in machine learning models [70]. Predicting cyberbullying successfully involves a set of extracted and constructed qualities [71]. In order to construct good cyberbullying prediction models, characteristics and their combinations must be taken into mind. Most studies on cyberbullying prediction [19], [38], [62], [72], [73] utilised machine learning approaches to create cyberbullying prediction models. Machine learning-based methods exhibit decent effectiveness in cyberbullying prediction [74]. Consequently, this study covers the building of cyberbullying prediction models based on machine learning.The machine learning field focuses on the study and deployment of computer algorithms that improve with experience [75], [76]. The objective of machine learning is to detect and characterise the patterns and relationships between data. The importance of researching massive data is in uncovering hidden information using deep learning from raw data [1]. Machine learning may be regarded as the adoption of computer models to increase machine performance by anticipating and characterising significant patterns in training data and the acquisition of knowledge from experience [77]. When this principle is applied to OSN material, the potential of machine learning resides in using past data to identify, predict, and analyse huge amounts of OSN data. For example, in supervised machine learning for classification application, classification is taught with the use of acceptable instances from a training dataset. In the testing phase, new data are input into the model, and instances are categorised to a specific class learned during the training stage. Then, classification performance is tested.This section addresses the most prevalent processes in the construction of cyberbullying prediction models for SM websites based on machine learning. The test comprises data collecting, feature engineering, feature selection, and machine learning approaches.

DATA COLLECTION

Data are key components of any machine learning-based prediction models. In the absence of knowledge or repercussions, data (even "Big Data") is meaningless. Selecting training and testing datasets based on data gathered from SM websites Prediction models that employ observed instances (labelled data) as input are known as supervised prediction models [78]. [78]. Most machine learning models are meant to generalise, but to do so, they require more than just samples from a training dataset [79]; they also need real-world data that has not been labelled. However, the quantity of data gathered is of little relevance; what matters is that the acquired data appropriately portrays activity on SM websites [80]–[82]. Cyberbullying prediction studies on SM websites have been classified into two categories: those that utilise keywords, such as words, phrases, or hashtags; and those that use user profiles (e.g. [19], [43]–[85]). (e.g., [38], [62], [70], [86]). To better understand how diverse data gathering strategies effect machine learning algorithms, the section on data collection has been introduced (related concerns) (related issues).

SPECIFIC RESEARCH AND TESTING

Any aspect of the work being observed may be quantified as a feature [87]. Machine learning algorithms may learn how to differentiate between various types of classes using feature vectors, which are the fundamental purpose of developing feature vectors [76]. In most machine learning models, feature engineering has a critical part in their success or failure. It is possible for a prediction to succeed or fail based on a number of circumstances. The most essential component is the training elements [78]. [61], [62], [72] are the most prevalent strategies used to create cyberbullying prediction models applying learning algorithms. The design of the input space (i.e., the attributes and their combinations provided as input to the classifier) is crucial in this circumstance.The most crucial step in developing an efficient machine learning classifier is to find a collection of discriminative qualities that may be employed as inputs to the classifier. Human-engineered observations may be utilised to construct feature sets based on the association between qualities and the occurrences of classes [76]. Recent studies [88]–[94] on cyberbullying, for example, established a relationship between factors like age, gender, and personality type with the chance of being bullied online. Because of this data (features), it is feasible to develop effective cyberbullying prediction models by adding them into the classifier's algorithm. Predictive models may be enhanced by adding features [76, 79]; this is a critical step. When constructing machine learning models for cyberbullying on social media, it is vital that researchers determine the most significant features of cyberbullying involvement on these platforms.Using cutting-edge research, features have been constructed to boost the accuracy of cyberbullying prediction. Using lexical syntactic traits, for example, it has been demonstrated that abusive language may be predicted with better accuracy than standard learning-based approaches. Using Myspace data, Dadvaret et al.

developed a gender-based approach for predicting cyberbullying based on profile information. In order to boost the classification performance of a classifier, the gender feature was utilised. Other studies [17], [61] included age and gender as factors, albeit they are based exclusively on the data that participants provide in their online profiles.In a number of study, profanity was exploited as a component in predicting cyberbullying. Proverbial phrases were also utilised as attributes for input into machine learning algorithms to detect bullying. [97, 98] The model's performance increases dramatically when profane phrases are added as features. For example, in an earlier research [70], features for input to machine learning included the volume and density of "bad" terms. Research discovered that a large frequency of "bad" phrases suggests cyberbullying in a correspondence. When a set of predetermined vulgar words was extended, researchers gave weights to develop bullying traits. [85] Feature input for a machine learning algorithm was concatenated with bag of words and latent semantic features.Reference In addition to basic models such as the bag of words (n-gram n = 1), [19] incorporated properties such as pronouns and skip grammes. These criteria, according to the authors, boosted the overall accuracy of categorisation. Studies [62] analysed textual cyberbullying related with Instagram comments and developed text characteristics comprising typical bagof-words features, picture comment counts, and post counts within less than an hour of releasing the image. The quantity of followers and likes, as well as shared media and aspects from picture material, such as image sorts, were added [62]. The overall classification performance was boosted by merging all of the criteria [62].

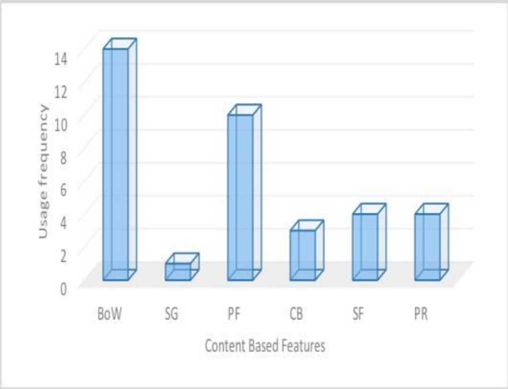

FIGURE 1. Depicting feature types used in cyberbullying prediction: Content-based features.

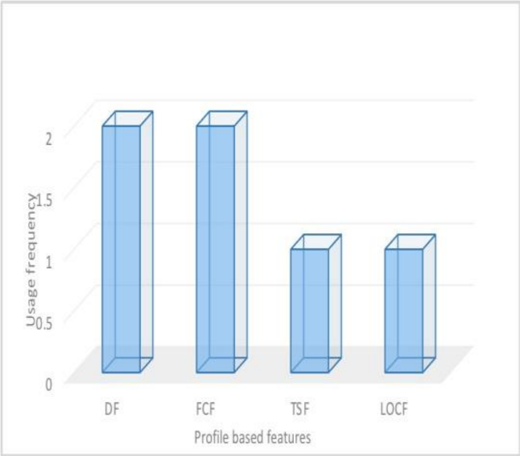

FIGURE 2. Depicting feature types used in cyberbullying prediction: Profile-based features.

When constructing the feature vector, a context-based approach is preferable than a list-based method. Cyberbullying's wide range and complexity, on the other hand, cast doubt on this generalisation. Sentiment analysis has been used in many research [68, 72, 96, 99] to help classifiers discriminate between cyberbullying and regular messages. Cyberbullying was thought to be linked to negative attitude in these research. Other studies have attempted to identify ways to reduce cyberbullying activities through the prediction of troll profiles, and one of these studies proposed an identification and association model for Twitter's troll profiles, believing that this is an important step toward predicting and stopping cyberbullying on SM websites [38]. New elements were recommended in this research to enhance the identification of authorship of postings and establish whether a profile is a troll or not. According to reference [99], the structure of SM websites (e.g., number of degrees, closeness, betweenness, and eigenvector centralities as well as the clustering coefficient) was combined with the characteristics of users (e.g., age and gender) and content (e.g., the number of articles per page) (e.g., length and sentiment of a post). The final machine learning accuracy is improved by combining these characteristics [99]. Table 1 illustrates a comparison of the many factors that have been utilised in cyberbullying prediction research. The accuracy of predictions may be impacted by this. Having a big number of traits that each correlate well with a certain lesson can help students learn. Thus, it is easy to see why so many experiments have strived for an abundance of characteristics. Textual cyberbullying is a common occurrence, and the input characteristics should reflect this. However, feature selection techniques should be used to assess the collection of features. To determine which characteristics are most likely to be relevant or irrelevant to classes, feature selection algorithms are used.

FEATURE SELECTION ALGORITHMS

Feature selection algorithms were rarely adopted in stateof-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.

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].

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

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

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.

$$I(Tr) = -\sum P_n \log_2 P_n /$$

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

ForattributeAttdatasets,theexpectedentropyiscalculate d as

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

The information gain of attribute *Att*datasets is

$$IG(Att) = I(Tr) - I(Att) \quad (3)$$

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:

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

where $\bar{x}$ and $\bar{y}$ 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

**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

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.

$$N = \frac{N[P(f, c_i)P(\bar{f}, \bar{c}_i) - P(f, \bar{c}_i)P(\bar{f}, c_i)]}{P(f)P(\bar{f})P(c_i)P(\bar{c}_i)}$$

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

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 ontheuseofBayestheoremwithstrong(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].

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.The number of examples (cases) in training data is set to N, and the number of attributes in the classifier is M.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.Therestoftheinstancesinthetrainingsetare used to approximate the error of the tree by forecasting their classes.

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.Alargenumberoftreesarethuscreated.Thesed ecision trees vote for the most popular class. These processes are called RFs [118].RF constructs a model that comprises a group of treestructured classifiers, in which each tree votes for the most popular class [118]. The most highly voted class is the selected as the output.

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 .... regarding the outputs for each example, the same procedure is applied recursively to each subset [120].

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].

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 wasusedintheconstructionofcyberbullyingpredictionmodels in [19] and [73].

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].

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:

ALGORITHMS FOR SELECTION IN COMPUTER LEARNING

The suggested features are sent into a machine learning algorithm, which is then trained on the results. Choosing the appropriate classifier for a particular dataset is tricky, though. To find the optimal machine learning algorithm for a certain dataset, it is necessary to evaluate a number of different algorithms. Machine learning algorithms may be narrowed down to a set of three factors. Machine learning for cyberbullying detection relies on a particular body of literature on the subject. The classifier's dominance may be restricted to a certain domain [134]. A machine learning algorithm may be chosen based on broad historical research and discoveries in the field of machine learning. Text mining literature review [135], [136] may also be utilised as a reference point. Machine learning methods may then be selected based on a comparison of extensive datasets [137]. These three factors may be used as a guide to limit down the selection of machine learning algorithms, however researchers need to test numerous machine learning algorithms in order to determine the ideal classifier for an accurate prediction model.

CLASS DISTRICTION IS IMBALANCED

When dealing with actual data, it is common for datasets to have an unbalanced number of instances of the normal class and an abnormal class. Real-world applications seldom include instances of abnormal classes, making it difficult to gather data on them. Among the uses of unbalanced data are fraud detection, instruction detection, and medical diagnostics. There are less cyberbullying posts than non-cyberbullying posts, and this assumption leads to an unbalanced distribution of posts in the dataset between the two classes: non-cyberbullying posts are much more numerous than cyberbully posts, and vice versa. As a result, the model may be unable to categorise the instances accurately. SMOTE [138] and weight adjustment (cost-sensitive methodology) [139] are two examples of approaches to this problem. Overfitting may arise when replicas of minority classes are introduced to the main dataset using the SMOTE approach [138]. Data from the minority class is used as an example and new synthetic classes are created to represent the rest of the population. The original dataset is then supplemented with these artificial classifications. The machine learning techniques are trained using the dataset that was produced. [139] The cost-sensitive method is used to regulate the imbalance class. To do this, a cost matrix must be created, and this matrix must identify the costs associated with false positives and false negatives that are discovered.

Characteristics of Human Data

Evaluation criteria like as accuracy, precision, recall, and AUC are often used [19], [38]. Choosing the right evaluation measure is critical. The choice is based on the characteristics of data that has been manually labelled. It's possible that selecting the wrong assessment measure will lead to greater performance. However, a study into how the machine learning model is assessed may give contradictory findings and may not accurately represent the gain in performance. Postings on cyberbullying, on the other hand, are often seen as outliers, while posts about non-cyberbullying are viewed as more typical. In general, the ratio of cyberbullying to non-cyberbullying is rather high. Non-cyberbullying messages often make up the majority. 1000 postings are manually classified as cyberbullying or non-cyberbullying, as an example. There are 900 postings that aren't cyberbullying, and there are another 100 that are. For example, if a machine learning classifier is unable to categorise a single post (0) as being cyberbullying, it is deemed ineffective. This classifier, on the other hand, has a high accuracy % if researchers choose accuracy as their primary assessment parameter, as is shown in the accuracy equation.There are no cyberbullying postings in the example, yet the classifier has a high accuracy rate. When deciding on an assessment measure, it's critical to understand the characteristics of data that has been manually labelled. It's possible that researchers will have to use AUC as their primary assessment measure when dealing with unbalanced data sets. AUC is more stable than other performance measures in class-imbalance scenarios [140]. Non-cyberbullying postings tend to outnumber cyberbullying ones in most datasets, which is a good representation of real-world data for machine learning algorithms. As a result, these algorithms' learning capabilities are unaffected by data skewness [73]. To prevent erroneous findings and properly assess the performance of machine learning algorithms, careful consideration should be given to the key evaluation measure.

POSSIBILITIES AND CONFLICTS

This section discusses the concerns and obstacles of identifying cyberbullying on social media using machine learning algorithms and models, while also pointing the way forward for future academics.

Characteristics of Human Data

Because of this, the examination of such vast datasets remains a matter of personal preference [141]. Developing human prediction systems necessitates engaging in processes fraught with subjectivity. For example, human bias may occur depending on the definition of cyberbullying and the criteria used to designate the language as cyberbullying content when constructing a manually labelled dataset to train a machine learning system to predict cyberbullying postings.Furthermore, subjectivity might arise throughout the feature engineering process of creating a collection of features (learning factors). As an example, a "data cleaning" procedure takes place at the pre-processing stage, whereby decisions regarding which characteristics are tallied and which are discarded are made. This is a subjective procedure by definition [141].Predicting human behaviour is critical, but also very difficult. A successful forecast of human behaviour requires that the patterns utilised to build a prediction model are also present in the data that will be fed into it. In order to maintain the model's context, the patterns should explicitly describe aspects that appear in both current and future data. There are a number of issues that arise while trying to comprehend and maintain a machine learning model based on huge data, which are not general and dynamic in nature. Big data context management is a difficult problem to solve, but it has been identified as an essential future focus [141].It is also important to remember that human behaviour is always changing. To keep the prediction model current, it's critical to track when and how online bullies alter their methods of cyberbullying. Human behaviour changes need dynamically updating the prediction model [1].

FACTOR CULTURE

For example, what was called cyberbullying a few years ago may no longer be deemed cyberbullying now because of the emergence of online social networks (OSNs). One of the characteristics of OSNs is that they are culturally diverse. Machine learning, on the other hand, is constantly influenced by the examples that are presented. There is still more work to be done in this area, and it will need the expertise of researchers from a variety of fields. Cross-discipline cooperation is essential for this task.

CHANGE IN LANGUAGE

Among the younger generation, language is rapidly evolving. Slang is always evolving and becoming a part of the lexicon. In this regard, researchers are urged to develop dynamic algorithms to identify new slang and abbreviations connected to cyberbullying activity on SM websites and to continually upgrading the training processes of machine learning algorithms by using newly presented phrases.

ASSESSING THE RISK OF CYBERCRIME

There needs to be an assessment of the severity of cyberbullying A cyberbully's impact is directly correlated to the extent and frequency of the bullying. For the prediction of various degrees of cyberbullying severity, it is necessary to conduct a thorough examination of social and psychological perceptions to define and classify the level of cyberbullying severity. To classify cyberbullying severity into different levels, rather than a binary classifier that only detects whether an instance is cyberbullying or not, efforts from various disciplines are required to define and identify the levels of severity. Then, related factors can be introduced and converted into features to build multi-classifier machine learning.

MACHINE LEARNING WITHOUT A MAN IN THE LOOP

Unsupervised learning is the norm in human development. The world's structure was found via observation, not by being given the names of all of its components. However, the success of supervised learning has eclipsed unsupervised machine learning. Many contemporary research depend on manually labelled data as the input for identifying classes, which may be the reason for this gap in literature. Because of this, detecting patterns between two classes through unsupervised grouping remains a challenge. To build unsupervised algorithms that can effectively find patterns in data, much investigation is necessary. Machine learning algorithms are unable to process large amounts of cyberbullying data.In recent years, academics in a wide range of sectors have been interested in deep learning. Deep learning is likely to have a major impact in the next few years in the field of natural language processing [142].The results of this study show that typical machine learning algorithms are unable to handle large amounts of data on their own. Traditional machine learning techniques have been proven ineffective due to the proliferation of large amounts of data. Big data on cyberbullying created by SM necessitates the use of innovative technologies in order to gather insights and aid in the formulation of intelligent judgments.Large amounts of data are being created at a rapid pace in terms of its variety (volume), variety (volume), verdict (value), truthfulness, complexity), and so on. The analysis of large amounts of social media data to detect cyberbullying practises requires the use of a variety of deep learning algorithms. It is possible to examine the enormous large data created by SM using several deep learning approaches and architectures, such as the generative adversarial network, deep belief network and convolutional neural network. In SM cyberbullying detection, these deep learning architectures have yet to be investigated.

CONCLUSIONS AND FUTURE DIRECTIONS ARE HERE

Machine learning techniques were used in this research to identify hostile conduct on SM websites. Data collection, feature engineering, creation of a cyberbullying detection model and assessment of the developed cyberbullying detection models were all examined in detail in our study. It also detailed a number of criteria used to identify cyberbullying on social networking platforms. Cyberbullying communications in online social networking sites may be classified using the most effective machine learning classifiers.It is a major addition to this research that the assessment criteria for machine learning algorithms can be defined so that the different methods can be compared. Machine learning approaches, particularly supervised learning, have been used to identify and summarise the most essential elements in cyberbullying detection. In order to get the area under the curve function for modelling cyberbullying, we employed accuracy, precision recall, and f-measure. Finally, the fundamental concerns and open research challenges were outlined and explored in detail.The development of detection algorithms that are both highly effective and extremely accurate in detecting cyberbullying requires a significant amount of study. It is our belief that this research will shed light on and shed new light on the identification of violent human behaviour, 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.PetersonandJ.Densley,''Issocialmediaagang?Towardaselection, 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.ShekokarandK.B.Kansara,''Securityagainstsybil attackinsocial 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. 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,andD.Smeed,''Thethreatsofsocialnetworking: 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.
22. Conf. Emerg. Netw. Exp. Technol., 2012, pp. 313–324.
23. 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.
24. 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.
25. J. W. Patchin and S. Hinduja, Words Wound: Delete Cyberbullying and Make Kindness Go Viral. Golden Valley, MN, USA:
26. Free Spirit Publishing, 2013.
27. 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.
28. 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.
29. 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.
30. 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.
31. 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.
32. 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.
33. M. McCord and M. Chuah, ''Spam detection on Twitter using traditional classifiers,'' in Autonomic and Trusted Computing. Berlin, Germany: Springer, 2011, pp. 175–186.
34. X. Chen, R. Chandramouli, and K. P. Subbalakshmi, ''Scam detection in Twitter,'' in Data Mining for Service. Berlin, Germany: Springer, 2014, pp. 133–150.
35. 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.
36. X. Zheng, Z. Zeng, Z. Chen, Y. Yu, and C. Rong, ''Detecting spammers on social networks,'' Neurocomputing, vol. 159, pp. 27–34, Jul. 2015.
37. 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.
38. *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.*
39. *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.*
40. *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 profilesinTwittersocialnetwork:Applicationtoarealcase ofcyberbullying,'' in Proc. Int. Joint Conf. SOCO-CISIS-ICEUTE. Cham, Switzerland: Springer, 2014, pp. 419–428.*
41. *Q. Huang, V. K. Singh, and P. K. Atrey, ''Cyber bullying detection using social and textual analysis,''inProc. 3rd Int. Workshop Socially-Aware Multimedia, 2014, pp. 3–6.*
42. *R. M. Kowalski, Cyberbullying: Bullying in the Digital Age. Hoboken, NJ, USA: Wiley, 2012.*
43. *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.*
44. *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.*
45. *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.*
46. *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.*
47. *H.Sampasa-Kanyinga,P.Roumeliotis,andH.Xu,''Associationsbetween cyberbullying and school bullying victimization and suicidal ideation, plans and attempts among Canadian schoolchildren,'' PLoS ONE, vol. 9, no. 7, 2014, Art. no. e102145.*
48. *S. Hinduja and J. W. Patchin, ''Bullying, cyberbullying, and suicide,'' Arch. Suicide Res., vol. 14, no. 3, pp. 206–221, 2010.*
49. *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.*
62.
50. *E. Whittaker and R. M. Kowalski, ''Cyberbullying via social media,'' J. School Violence, vol. 14, no. 1, pp. 11–29, 2015.*
51. *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.*
52. *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.*
53. *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.*
54. *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.*
55. *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.*
56. *Q. Li, ''New bottle but old wine: A research of cyberbullying in schools,'' Comput. Hum. Behav., vol. 23, no. 4, pp. 1777–1791, 2007.*
57. *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.*
58. *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.*
59. *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.*
60. *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*
61. *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.*