IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 15, 2021, accepted June 8, 2021, date of publication June 15, 2021, date of current version June 28, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3089515*

# Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review

**NANLIR SALLAU MULLAH**[1,2], (Member, IEEE), AND **WAN MOHD NAZMEE WAN ZAINON**[1]

[1]School of Computer Sciences, Universiti Sains Malaysia, Penang 11800, Malaysia
[2]Federal College of Education Pankshin, PMB1027 Pankshin, Plateau State, Nigeria

Corresponding authors: Nanlir Sallau Mullah (mullakns@gmail.com) and Wan Mohd Nazmee Wan Zainon (nazmee@usm.my)

**ABSTRACT** The aim of this paper is to review machine learning (ML) algorithms and techniques for hate speech detection in social media (SM). Hate speech problem is normally model as a text classification task. In this study, we examined the basic baseline components of hate speech classification using ML algorithms. There are five basic baseline components – data collection and exploration, feature extraction, dimensionality reduction, classifier selection and training, and model evaluation, were reviewed. There have been improvements in ML algorithms that were employed for hate speech detection over time. New datasets and different performance metrics have been proposed in the literature. To keep the researchers informed regarding these trends in the automatic detection of hate speech, it calls for a comprehensive and an updated state-of-the-art. The contributions of this study are three-fold. First to equip the readers with the necessary information on the critical steps involved in hate speech detection using ML algorithms. Secondly, the weaknesses and strengths of each method is critically evaluated to guide researchers in the algorithm choice dilemma. Lastly, some research gaps and open challenges were identified. The different variants of ML techniques were reviewed which include classical ML, ensemble approach and deep learning methods. Researchers and professionals alike will benefit immensely from this study.

**INDEX TERMS** Text classification, cyber hate, deep learning, ensemble technique, machine learning, social media networks.

## I. INTRODUCTION

Social media networks (SMNs) are the fastest means of communication as messages are sent and received almost instantaneously [1], [2]. SMNs are the primary media for perpetrating hate speeches nowadays. In line with this, cyber-hate crime has grown significantly in the last few decades [3]. More researches are being conducted to curb with the rising cases of hate speeches in social media (SM). Different calls have been made to SM providers to filter each comment before allowing it into the public domain [4], [5].

The impacts of hate crimes are already overwhelming due to widespread adoption of SM [6] and the anonymity enjoyed by the online users [7]. In this era of big data, it is time-consuming and difficult to manually process and classify massive quantities of text data. Besides, the precision of the categorization of manual text can easily be influenced by human factors, such as exhaustion and competence. To achieve more accurate and less subjective results, it is beneficial to use machine learning (ML) approaches to automate the text classification processes [6]. There have been significant advancements in ML techniques from classical ML, ensemble and deep learning (DL) techniques for hate speech detection. Due to the unprecedented advancement in natural language processing (NLP), several machine learning methods have achieved superior outcomes [8].

To be able to improve classification of SM texts as hate speech or non-hate speech, researchers and practitioners require an updated understanding of machine learning methodologies, which is fast evolving. Considerable effort has been spent on creating new and effective features that better capture hate speech on SM [9]–[11]. Slangs and new vocabularies are also constantly evolving in the SM space. New and updated datasets are also available across different regions of the world. To bridge the gap, there is a need to review the literature and keep professionals, old and new researchers in the know of the currents developments in this research area. On this note, this review becomes necessary to be conducted.

The remaining parts of this article are structured in the following ways: Motivation and Related Works are presented in section II. Section III covers the methodology. The

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

concept of hate speech and hate speech modelling is covered in section IV. Hate speech classification, contribution and limitations of past works, open challenges in hate speech detection, limitation of the study and conclusion are covered in section V, VI, VII, VIII and IX respectively.

## II. MOTIVATION AND RELATED WORKS

### A. MOTIVATION

The cases of hate speeches have become rampant due to the SM adoption by a large population. Researches have shown that hate speeches can influence political discourse and can change the narrative negatively [12] , [13]. It is of great importance to police the SMNs to allow democracy to take it natural cause without undue influence through hate speech spread.

It is also obvious that countries where their democracy is still at the infant stages are more vulnerable in the face of hate speeches than those with matured democracy. Therefore, developing a hate speech detection system can help in keeping countries in mutual coexistence.

Committing cyber hate requires just a smartphone, internet connection and a person with a corrupt mind. The hate speech post can be escalated to every nooks and cranny in a matter of seconds. A geographical boundary is not a limitation in posting and spreading hate speeches on SMNs. Therefore, developing an effective hate speech detection on SM is of great significance. There is nothing the targeted person or group can do to stop the spread of this offensive post [14]. To a reasonable extent now, SM is an integral part of our daily lives [15].

It is necessary to fight the systematic racism rooted in almost all societies around the globe. JPMorgan Chase has promised to commit USD30 billion over the next five years to advance racial equity[1] [16]. JPMorgan Chase Chairman and CEO Jamie Dimon, said they need to do more to truncate systems that have propagated racism and widespread economic inequality, especially for Black and Latino people. Following the police shootings of George Floyd and Breonna Taylor, there has been an increase in philanthropic giving for fighting racism as a variant of hate speech [16]. This study is also a timely contribution in reducing hate speech on social media.

### B. RELATED WORKS

Abusive messages in social media is a complex phenomenon with a broad range of overlapping modes and goals [17]. Cyberbullying and hate speech are typical examples of abusive languages that researchers have put more interest in the past few decades due to their negative impacts in our societies. Several research have been conducted to automatically detect these undesirable messages among other messages in social media.

The automatic detection of hate speech using machine learning approaches is relatively new, and there are very limited review papers on techniques for automatic hate speech detection [18]. The recent and related survey papers available on review of hate speech detection methods during this research work were few. The following were the available traditional literature review related to automatic detection of hate speech using MLA: [19], [20]

ML algorithms have contributed immensely in hate speech detection and SM content analysis generally [15]. Offensive comments such as HS and cyberbullying are the most researched areas in NLP in the past few decades [21]. ML algorithms have been of great help in this direction in terms of SM data analysis for the identification and classification of offensive comments [22]. The advances in ML algorithms researches have made significant impacts in many fields of endeavour which led to some important tools and models for analysing a large amount of data in real-world problems like SMNs content analysis [23].

In this survey conducted by [20], the authors presented a brief review on eight hate speech detection techniques and approaches. These eight techniques include TF-IDF, dictionaries, N-gram, sentiment analyses, template-based approach, part of speech, Bag of the word, and rule-based approach. The limitation of the review is that techniques such as deep learning and ensemble approach were not considered in their work.

In [19], the authors offered a brief, and critical analysis of the areas of automated hate speech detection in natural language processing. The authors also analysed the features for hate speech detection in literature which includes: simple surface features, word generalization, sentiment analysis, lexical resources, linguistic features, knowledge-based features, meta-information and multimodal information.

The limitation of these two reviews is that techniques such as deep learning and ensemble approach are not considered in their work. The most significant step in text classification pipeline is selection of the best classifier [8]. Therefore, the need to review all techniques is of essence. We intent to make this selection phase easier for researchers by reviewing more algorithms than the previous review work have covered. In this case, we reviewed techniques like deep learning, ensemble learning among others that have been employed for the automatic detection of hate speech in social media.

Posters of hate speeches usually attack their targets using the following attributes: Religion, Race, political affiliation, gender, marital status, ethnicity, health status, disability and nationality [24]. The data generated by SM sites are increasing in the geometrical proportion daily called big data [15]. About 7.7 billion population of the world [25], [26], the following approximate population are actively connected on one social site or the other [27]–[29], as shown in figure 1.

The research involving this large population, and to understand the trend of the behaviour of humans is of paramount importance. A problem that can be caused by a large population such as this cannot be ignored.

[1] https://www.jpmorganchase.com/news-stories/jpmc-commits-30-billion-to-advance-racial-equity

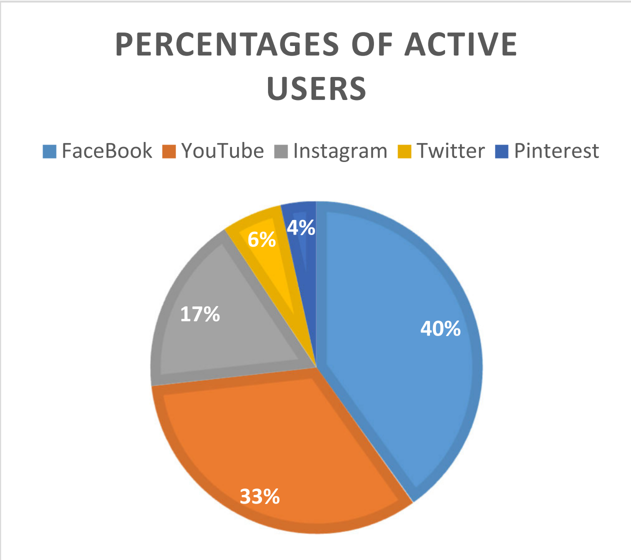


FIGURE 1. Active users on social media.

## III. METHODOLOGY

The methodology used for this work is explained as follows. The following databases were mainly used to get the required articles for this review work: IEEE Explore, ACM, ScienceDirect, Scopus and Universiti Sains Malaysia databases. These databases were used because of their reputation and also they are subscribed by Universiti Sains Malaysia Library. The researchers limit the articles search to a span of ten (10) years (2010-2020) for the review work. Key terms or phrases used in the search retrieval includes hate speech detection, offensive comments, aggressive comments, cyberbullying, profanity and toxic comments on SM.

The filter tools available in each database were used to filter the articles. For instance, the subject was restricted to computer science, engineering, and mathematics. In this case, only the most relevant were downloaded after all filter tools have been employed. The second phase involves going through the abstract of each article to apply the inclusion or exclusion criteria. Those papers that passed the inclusion test, were sorted according to their years of publication. The first inclusion criterion is that the paper must have addressed issues related to offensive comments (hate speech, cyberbullying, aggressive comments, toxic comments, etc.) on SM. Two sections of each paper were used for this purpose: the title and the abstract.

## IV. THE CONCEPT OF HATE SPEECH AND HATE SPEECH MODELLING

### A. THE CONCEPT OF HATE SPEECH

Hate speech refers to any kind of communication in speech, writing or behaviour, which attacks or uses pejorative or discriminatory language regarding a person or a group based on some sensitive information or protected characteristics [5], [30]. These protected characteristics include religion, ethnicity, nationality, marital status, health status, race, colour, disability, sexual orientation, descent, gender or other identity factors [31]. Hate speech is a widespread phenomenon and has become an accepted reality as a common enemy of all law-abiding citizens across the world. This is a dangerous and illegal act that needs to be discouraged! Most of the hate speech messages on SM are constructed through texts [32]. However, images and sounds are also used in the dissemination of hate speeches [32] . Therefore, any attempt to address this problem through Computer perspective, text classification is the best bet.

There is no universally accepted definition of hate speech, no consensus agreement on an individual definition [33]. It has been observed that a clearer and precise definition of hate speech can simplify the annotators work and consequently increase the annotators' agreement rate [34]. Although, it can be difficult in some countries to differentiate between appropriate speech and hate speech. Hence, giving a precise and universal definition of hate speech become more difficult and complicated. For example, there is a thin line between hate speech and normal speech under the First Amendment in the US. However, any speech that contributes to a criminal act is punishable as part of a hate crime. The debate on what can be classified as hate speech is not new, but there are conscious and renewed efforts as the world experience the Black Lives Matter (BLM) movement across the world. The BLM movement came up after the death of George Floyd.

Beside hate speech, there are other abusive online behaviours which are worthy of clarification, such as cyberbullying. Cyberbullying as a kind of cyber harassment [35] means repetitive hostile behaviour through SM in an attempt to deliberately and consistently threaten or hurt individuals who cannot defend themselves easily [36], [37] and is common among youth [38], [39]. Cyber-hate or Hate speech and Cyberbullying are all different forms of abusive online behaviour [17], [36]. Cyberbullying can be considered as Hate speech when sensitive or protected feature of a victim is the target of the attack. Hate speech is distinguished from cyber-bullying such that hate speech will affect not just a person but does have consequences for the entire group or society [18]. Hate speech is a complicated and multi-faceted concept that has been difficult to understand, by both human beings and computer systems [40].

### B. HATE SPEECH MODELLING

Hate speech detection problem is normally formulated as a text classification task. The initial pipeline input consists of some raw texts data. Generally, text datasets can be modelled mathematically as $D = \{a_1, a_2, a_3, \ldots, a_n\}$ where D is a sequence of text documents, $a_i$ is a data point having N sequences of sentences, in which a sentence includes $w_N$ words with $p_w$ letters [8]. A unique point is classified with a label value from a set of v different discrete value indices [41].

## V. HATE SPEECH CLASSIFICATION

Over the past few decades, text classification has been researched extensively and used in many real life applications such as hate speech detection. More researchers are now interested in developing applications that leverage text classification methods, especially with recent advances in NLP and text mining. Generally, hate speech classification leveraging ML can be grouped into five phases: Data collection and exploration, feature extraction, dimensionality reduction, classifiers selection and evaluations as summarized in Figure 2.

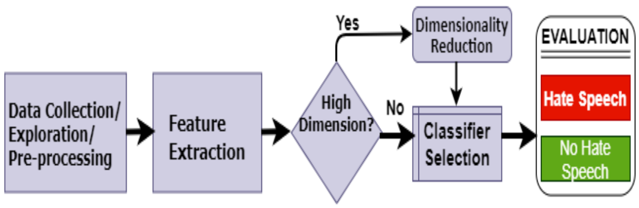


**FIGURE 2.** Hate speech detection components using ML.

### A. DATA COLLECTION AND EXPLORATION

This is a stage where the researcher will make a decision pertaining to how and where data will be obtained for the training of the machine learning algorithm of choice. A researcher may be lucky to get published dataset or unlucky and have to create a new dataset from the scratch. There are two things to consider whether a published dataset will be used or new one created – availability and relevancy [42].

Dataset may not be available at all or completely obsolete. In this case, we are left with the option of creating new dataset or update the old one. Creating a new dataset is a laborious and expensive undertaking but in most cases, it worth the time and cost.

The relevancy of the available dataset is central to the choice of the data set to use for building any predictive model. Before a dataset is labelled, certain criteria are spelt out based on the nature of the problem to be solved. If the current research goal is the same with the one the dataset was created for, then it can be easily be adopted as seen in [43]–[46]. However, new dataset will become necessary when no old and relevant dataset is available.

### B. FEATURE EXTRACTION

Texts generally are unstructured data. However, all ML algorithms use mathematical modelling as an integral part of the algorithm, therefore, the unstructured nature of the texts data must be converted into structured feature space [10]. The noise such as unnecessary numbers, common words, non-English words in the dataset must be gotten rid of. When the dataset is cleaned, vectorization methods can be used to convert the dataset into a vector space.

### C. DIMENSIONALITY REDUCTION

In this era of big data, the volume of data generated is increasing per second, especially in the SM space. It is also true that finding a meaningful trend in this huge data is becoming very difficult due to the presence of less important data [47], [48]. These irrelevant data are even more in number than the important ones [49]. This makes the data generally sparse and unevenly distributed over the search space and also referred to as high dimensional data. The difficulty of identifying trends in this our big data era due to the high dimensionality of data is referred to as the curse of dimensionality [50]. To use this dataset for training a model, most of the unimportant data must be reduced to the barest minimum for maximum performance of the classifier.

This problem is handled through technique called dimensionality reduction. Every ML experts strive to clean the data of any noise and remove some features that will not add learning value to the model. In an attempt to do this, other problems can set in like overfitting and data leakage. Overfitting occurs when data is too few and the classifier learns too little as well and when faced with unknown data, it performs poorly. Data leakage occur when in the process of splitting the few data available for cross-validation, and it happens that the training data and testing data contains some data in common. This will make the accuracy very high but when expose to a new dataset, the classifier will fail woefully. This problem can be solved through obtaining a critical dimension of the data set.

A critical dimension of a data set is the minimum feature set required to train a classifier and capable of predicting with reasonably high accuracy [47], [48]. Critical dimension usually guides researcher from over reducing the features in the features space which may lead to overfitting. When the dimensionality reduction technique has been applied, the classifier should able to learn enough using the reduced features and perform the clustering or classification task optimally.

### D. HATE SPEECH CLASSIFIER SELECTION

Hate speech problem is normally model as a text classification task. There are different classifiers out there to use for hate speech classification problem. One of the most significant steps in hate speech identification pipeline is selecting the optimal classifier. To accomplish this, there is a need to have a complete conceptual understanding of each hate speech classifier to guide algorithm choice. Machine learning is generally classified into classical method, Ensemble approach and deep learning method [51]. The key aspect that we are concern on in this paper is advances made so far in these methods. The comparison of some related techniques deployed in recent time is shown in Table 1.

#### 1) CLASSICAL MACHINE LEARNING

This approach is also called shallow method. This method relies on manually or automatically coded dataset that can be used for training purposes. This labelled dataset is used to train the learning algorithms to produce a model which can be used for detecting and classifying text as hate speech or non-hate. Examples include support vector machines (SVM), Naive Bayes (NB), Logistic Regress (LR), Decision

**TABLE 1.** Comparison of related techniques for hate speech dection.

| Author | Classifier | Novelty | Feature Extraction | Evaluation Metrics |
|---|---|---|---|---|
| [52] | NB, RF, LG, DT, SVM, DL | Improvement on islamophobia detection | Word embedding | Accuracy, precision, recall and F1 |
| [53] | DL | HS in Context | embedding | Accuracy, Recall, Precision, F1-score |
| [54] | Ensemble method | Multi-tier meta-learning model | character n-gram and word n-gram | Precision, Recall and F1-score |
| [45] | GRU | A new study on the Amharic language | Word2Vec | Accuracy, ROC, AUC |
| [55] | SVM, NB, DT, RF | To detect Arabic context-based HS | BoW and TF-IDF | Accuracy, precision, recall, G-mean |
| [56] | NB, LR, SVM, KNN, DT, RF | Addresses Code-switch | TF-IDF | Confusion matrix |
| [51] | LR and LSTM | Multi-lingual aspect analysis of HS | BoW | F1-score |
| [57] | RF | Improved RF for HS detection | Count vectors | F1-score, precision, recall |
| [58] | Lexicon, RNN | The building of Arabic dataset | N-gam, embedding | F1-score, precision, recall, AUROC |
| [59] | SVM, NB & RF | Emotional Analysis | N-gram | Precision and Recall |
| [3] | RF, SVM, J48graft | Combination 3 different dataset which gives a wider coverage | Unigrams | Precision, Recall, F1 |
| [60] | n-Gram word | Identifying cyber hate | BoW | Precision, Recall, F1 |

Trees (DT), K-Nearest neighbour (KNN), etc. The commonly used ML algorithms for hate speech detection are summarized in Figure 3.

From Figure 3, SVM has the highest number of usage by researchers to classify SM data as hate speech or non-hate speech. Random forest is the second in the ranking, logistic regression and naïve Bayes are used considerably well too.

### 2) ENSEMBLE APPROACH

The ensemble approach is simply applying the wisdom of the crowd. In other words, the aggregate predictions of many classifiers are always better than the best single classifier [61]. The ensemble technique was designed to address the weaknesses of the various individual ML algorithm and consolidate their strengths [62]. It is evident that each model has its share of pitfalls; therefore, no model is perfect. Though the ensemble methods try to add up the advantages of other models together to give a better performance than any single model can offer [63]. Statistically speaking, combining two or more ML algorithms can generally reduce their variance and significantly improve their learning capabilities [64]. There are different types of ensemble techniques which include; random forest, bagging approach and boosting method. Each of these methods has its strength and weaknesses in handling hate speech task as summarized in Table 2.

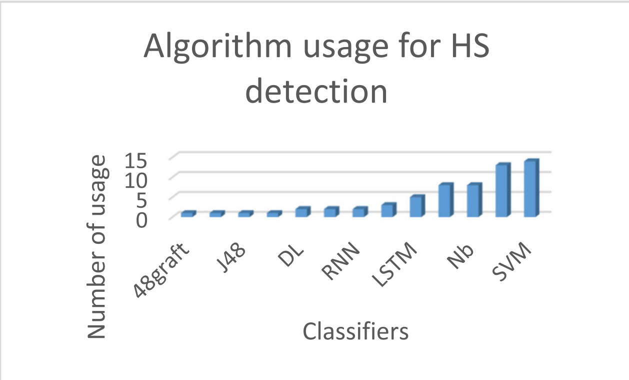


**FIGURE 3.** Classifier usage rate.

### 3) DEEP LEARNING METHOD

Some texts datasets are very large and not linearly separable, therefore, classical ML cannot analyse it effectively. Data that are not linearly separable are simply nonlinear data that the hyperplane cannot be easily drawn. To solve this problem of predicting meaningful trends in linearly non-separable data, the DL algorithm was proposed. DL is simply an extension of ML algorithm called artificial neural network (ANN) [65]. The deepness depends largely on the complexity of the problem at hand. Image processing task for instance, usually requires deeper layers than SM text prediction tasks [66], [67]. The attention of researchers has been attracted to Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) because they capture sentence semantics better. CNNs particularly have proven to be effective in capturing semantics and syntax of words in contents analysis [68].

Different variants of deep learning have been applied to detecting hate speech in social media. [69] applied CNN and two variants of RNN, which are Long Short Term Memory (LSTM), and Gated recurrent units (GRUs) to solve Task6 of SemEval-2019, which requires participants to identify and classify offensive text in SM. In this [69], the researchers also experimented two approaches proposed by [70]; LSTM-CNN and CNN-LSTM models. In the end, [69] concluded that BiLSTM-CNN gave a better F1-score. Another research conducted on hate speech detection and three deep neural networks (DNN) was applied [71]. In this research, the following variants of DNN were used; FastText, CNNs and LSTMs. [71] research outperformed the state-of-the-art by approximately 18 points better.

The obvious difference between DL and ML is that DL requires large dataset to learn reasonably, while ML require less to learn as can be seen or described by the learning graph in Figure 4.

The red line indicates the learning curve of deep learning algorithms. The curve keeps growing alone the performance-axis (vertical axis) with increase in data, this growth represents the performance of the algorithm. That means the more data, the better the performance of deep

**TABLE 2.** Weaknesses and strengths of ensemble techniques.

| Ensemble Technique | Weakness | Strengths |
|---|---|---|
| Random forest | Requires ample memory space when handling large dataset<br><br>Slow to produce predictions once trained<br><br>The complexity of the prediction step is directly proportional to the number of trees in the forest.<br><br>It is necessary to choose the number of trees in the forest<br><br>Relatively difficult to interpret | Proven to be the most accurate ML algorithm<br><br>Training speed is faster and easy to make a parallel method<br><br>Relatively simple to implement<br><br>For unbalanced data sets, it balances the error.<br><br>If a large part of the features is lost, accuracy can still be maintained.<br><br>Can handle data with high dimensionality and overfitting problem<br><br>Can automatically select the best features required for the learning process |
| Bagging | Not good in the case of bias or under-fitting.<br><br>The value with the highest and lowest result, which can have a significant difference and have an average result, is typically overlooked. | It has been shown to reduce the variance in a classification task<br><br>This creates an atmosphere by the use of N learners of the same size on the same algorithm to deal with variance. |
| Boosting | High computation time<br>Sensitive to noise<br><br>Susceptible to outliers although the errors in the predecessors must be corrected by each classification algorithm.<br><br>It is almost not possible to scale up.<br><br>It can ignore overfitting in the data set.<br><br>It increases the complexity of the classification.<br><br>Time and computation can be a bit expensive. | Decreases the variance of the classification as well as its bias.<br><br>Can yield more reliable outcomes for classification.<br><br>A record of net errors is kept at every point of its results<br><br>This performs the weighting of the larger sampling accuracy and smaller sampling accuracy and then provides the cumulative performance.<br><br>Helps when dealing with bias or under-fitting in the data set. |

learning algorithms. On the other hand, traditional machine learning which is represented by the blue line, indicates that the algorithm will certainly stop learning even if the data con-

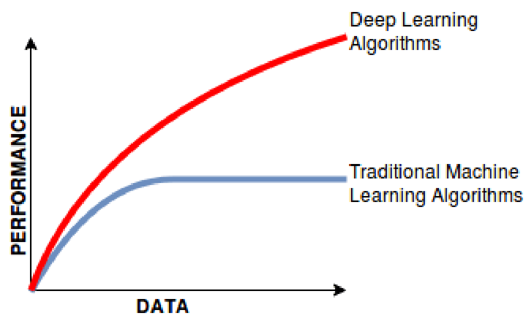


**FIGURE 4.** Traditional ML vs DL learning curve.[3]

tinue to increase. The horizontal blue line means no further learning is taking place.

Large number of previous studies carried out on automatic hate speech detection mostly focused on traditional machine learning for detection of various forms of hate speech in the social media. The data generated on social media is at an exponential rate, hence very large [72]. This call for the use of deep learning to solve the problem. There are very few papers on deep learning for hate speech detection. TABLE 3 shows some comparison of deep learning for hate speech detection.

**TABLE 3.** Comparison of deep learning for hate speech detection.

| Author | Aim of the Study | Futures Extraction method | Deep Learning Algorithm | Evaluation metric |
|---|---|---|---|---|
| [73] | To solve discriminatory problem | word embedding | CNN | std deviations = 0.84 |
| [17] | To identify hate speech in Arabic Tweets | character n-gram and CBOW | CNN and RNN | Pr= 0.81, Rc= 0.78, A = 83, F1 = 0.79, AUC = 0.89 |
| [74] | To improve the performance | CBOW and Continuous Skip-gram | CNN, LSTM, CNN+GRU | F1 = 93.35 |
| [71] | To classify a tweet as racist, sexist or neither | Char n-grams, TFIDF, BoWV | CNN and LSTM | Pr= 0.93, Rc= 0.93, F1 = 0.93 |
| [43] | Detection and explanation of hate speech on SM | NA | Deep LSTM | A = 90.82, Pr= 83.82, Rc = 84.23 |

### *E. PERFORMANCE EVALUATION METRICS FOR HATE SPEECH DETECTION*

Performance evaluation is a research problem across all disciplines, which is usually carried out using performance evaluation metrics. Performance evaluation metrics are logical-mathematical constructs obtain by the difference between the actual values and the predicted values [75].

Performance evaluation of hate speech detection models typically makes use of the classic precision, recall and F1-score metrics. These are mostly used because of the unbalanced nature of the hate speech dataset. For any balanced dataset, accuracy is the best option. The Precision, recall, accuracy and F1-score evaluation metrics are clearly explained in [15], [65], [76].

Suppose our model was trained to classify tweet as hate speech and non-hate speech. For instance, we have a set of 20 tweets containing 5 tweets as hate speech and 15 as non-hate. The model was able to identify 6 tweets as hate speech. Of the 6 tweets identified, 4 were actually hate speech (true positives) and 2 were non-hate speech (false positive). The model misclassified 2 tweets (false negative) which were hate speech and 13 tweets were accurately excluded as non-hate (true negatives).

#### 1) PRECISION ($P_r$)

Precision is the ratio of true positive and total predictions. The following researchers made use of precision to evaluate their model performance; [43], [52]–[54].

This can be represented mathematically as:

$$P_r = \frac{TP}{TP + FP} \tag{1}$$

$P_r$ is a short for precision for the purpose of this study. Precision simply means a fraction of positive classifications that was correctly identified by the model [77]. For example, the proportion of actual positives that were identified correctly from the example above is 4. Then the model precision is 4/6 (true positives / all positives) = 0.67.

*TP* is a short for true positive. From the scenario above, TP is 4. Out of 5 hate speech tweets, the model was able to correctly identify 4 as hate speech.

*FP* means false positive. This refers to non-hate speech tweets that were classified as hate speech. From the scenario above, 2 tweets were missed classified as hate speech tweets and in the real sense, they were non-hate speech tweets.

#### 2) RECALL ($R_c$)

$R_c$ is the ratio of the number of correct predictions and all correct observation in the sample space. [55], [57], [78] and [79] made use of recall for their evaluation. Mathematically:

$$R_c = \frac{TP}{TP + FN} \tag{2}$$

$R_c$ stands for Recall in this paper. This means the proportion of real positives that were established correctly. From the scenario, recall is 4/5 (true positives / all positives) = 0.8. This means the model was able to correctly identify 80% of the hate tweets.

*FN* stands for false negative for the purpose of this study. This refers to those hate speech tweets that were not identified by the model as hate speech. The model considered them as non-hate while they were hate tweets in the real sense. In the example above, only one tweet was misclassified as non-hate and was actually hate speech.

#### 3) F-MEASURE

F-measure (F) or F1-score (F) is simply the weighted harmonic mean (whm) of precision and recall. This evaluation metric is normally employed when the dataset is unbalanced. It was employed to evaluate performance of hate speech prediction model in [51], [52], and [57]. Mathematically:

$$F = 2 * \frac{P_r * R_c}{P_r + R_c} \tag{3}$$

$F$ is short for F-measure or F1-score and is used to test the model's performance with an imbalanced class distribution. In most real-life text classification tasks, imbalanced class distribution occurs and hence F1-score is a smarter metric to test a model [51]. From example above, $F = 2(0.67*0.8)/(0.67 + 0.8) = 1.072/1.47 = 0.72$. This simply means that the F1-measure of the model is 72%.

#### 4) ACCURACY (A)

Accuracy is the ratio of correct prediction and total observations. Accuracy of a model is considered best if and only if we have symmetric dataset in which the value of FP and FN are almost equal for the two-class problem. Accuracy is not the best option in multiple and imbalanced data sets, hence, other evaluation parameters may be considered, like F1-score. In the following researches, [45], [52], and [80], accuracy was used. Mathematically, accuracy (A) can be expressed as:

$$A = \frac{TP + TN}{TP + FP + FN + TN} \tag{4}$$

## VI. CONTRIBUTION AND LIMITATION OF THE STATE-OF-THE-ARTS

Table 4 presents the contribution and limitations based on the article that has been reviewed.

Form Table 4, the following gaps are obvious for further research. Numeric symbols and special characters which may connotes or convey hate speech messages and were ignore in all papers reviewed. A comprehensive coding guide benchmark is always necessary to guides the annotators. More research is required to handle hate speech messages that are contextual in nature.

## VII. OPEN CHALLENGES IN HATE SPEECH DETECTION

These are some of the hurdles associated with the detection of hate speech in the SM through leveraging ML algorithms. These challenges come in different ways ranging from the definition, dataset collection and annotation, cultural variation and other associated problems.

### A. DATASET AND HATE SPEECH DETECTION CHALLENGE

The first fundamental problem is the availability of hate speech dataset across different regions of the world. To carry out analysis on SMNs, a large dataset is significant [60], [84] has observed that there is an urgent need to take the campaign of hate speech prevention to other non-western parts of the world. This means that culture and tradition play a significant role in hate speech detection efforts. Table 5 shows the dataset availability across different regions of the world.

### B. DATA SPARSITY CHALLENGE

The second problem is the sparsity of the dataset. For example, on Twitter, only 140 characters are allowed per post [87]. In this case, the information in a given tweet may not be sufficient to generalize on a particular post. This is a common problem across all short messaging text mining task.

### C. UNBALANCED DATASET CHALLENGE

The problem of imbalance class distribution nature of dataset in hate speech detection is a commonplace, as this occur naturally in most real life problems [51]. In most cases, the normal (non-hate) post is more than the abnormal (hateful) posts [88]. This will lead to bias learning as the algorithm will learn more on the majority class (non-hate) data than minority class (hate speech) data.

### D. CULTURAL VARIATION

Cultural variations directly affect the definition of hate speech or what constitute hate speech varies with culture and tradition. What is considered in the US a normal speech can be seen as hate speech in Nigeria for example. The culture and tradition of people play a great role in the classification of speech as offensive or non-offensive. Experts have recommended that for the SM providers to holistically address the hate speech problem on their platforms, the non-western regions of the world must be considered for hate speech related researches [13].

### E. PANDEMIC OR NATURAL DISASTER

Pandemic or natural disaster victims can be stereotyped. The typical example is the COVID19 pandemic, where Chinese have been stereotyped in many places across the globe. See the typical stereotyping tweet by former President Trump in Figure 5.

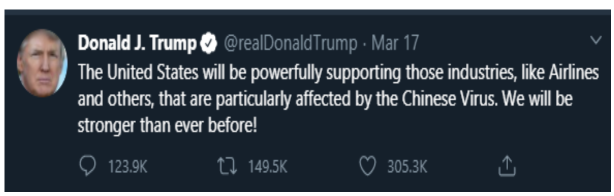


**FIGURE 5.** President trump tweet.

From the tweet in Figure 5, Trump described the COVID19 as Chinese Virus. This description did not go

**TABLE 4. Contributions and limitations of related works.**

| Related Work | Dataset Source | Contributions | Limitations |
|---|---|---|---|
| [52] | Twitter | • The new dataset was created<br>• Guideline for annotation was derived by experts<br>• Clear definition of Islamophobia was stated to guide annotators<br>• The inter-coder agreement was high enough, 89.9% | • The data collected was restricted to those following the key politicians in the UK; it limits the spread.<br>• The data collection was not heterogeneous due to restrictions<br>• Concentrate on Islamophobia only; another hate-related aspect was not considered<br>• Words context do not matter<br>• Numeric symbols were removed as part of pre-processing, which could be valuable information |
| [53] | Twitter | • Good heterogeneous coverage of tweets<br>• Most hate variables were considered | • Proper annotation guideline is required, as annotation was done based on overall perceived meaning.<br>• Sensitive hate-related areas such as health status, marital status, transgender were not considered<br>• Special characters and numeric symbols were removed as part of pre-processing |
| [54] | Twitter | • Good general coverage of heterogeneous society<br>• Annotators which were experts in South African politics were trained before labelling the dataset | • The dataset does not take care of other languages in South Africa, except for code-mixed<br>• Numeric symbols were removed as part of pre-processing<br>• Using even number (i.e. 2) annotators was not a good idea because, for the unclear post, one can say hate, and the other can say non-hate. That can be a serious problem. |
| [45] | Facebook | • Good examples of annotators instruction<br>• Good coverage of hate variables | • The published dataset was used and also inherited issues associated with the dataset<br>• Code-mixed data post was not considered<br>• Numbers which some may covey useful meanings was not considered<br>• Transgender and marital status was not put into consideration |
| [81] | Twitter | • A good study of cyber-hate in other languages besides English<br>• Very good cross-validation of up to 10 | • Only Spanish text was studied<br>• Code-mixed texts were removed which may lead to loss of important information<br>• Numeric symbols and pictures and emojis were not considered |
| [82] | Facebook | • Good examples of annotators instruction<br>• Good coverage of hate variables<br>• Cohen's kappaκ value was computed to test the inter-coder agreement | • Code-mixed data posts were not considered<br>• Numeric symbols which some may covey useful meanings was removed<br>• The transgender and marital status were not put into consideration |
| [9] | Twitter | • A comprehensive data was collected, stating from lexicon | • Numeric symbols were removed as part of pre-processing |

**TABLE 4.** *(Continued.)* **Contributions and limitations of related works.**

| | | | |
|---|---|---|---|
| | | creation<br>• The context was put into consideration and not just considering negative words.<br>• Proper definition and explanation were given to guide the annotators | • No comprehensive guideline to help the annotators |
| [83] | Twitter | • Code-switching among different language speakers was considered<br>• Multi-lingual and multi-aspects were both considered in the studies | • Numeric symbols and special characters were ignored<br>• The contextual issue in texts was not addressed |

**TABLE 5.** **Geographical distribution of cyber-hate dataset and availability comparison.**

| Reference | Domain | SM Source | Availability | Dataset Source | Origin (Country) |
|---|---|---|---|---|---|
| [43] | General | Twitter | Available | Adopted [9] | USA |
| [52] | Specific (Politics) | Twitter | Available | New | UK |
| [44] | General | Twitter | Available | Adopted [9] | USA |
| [53] | General | Twitter | Not available | New | Jordan |
| [54] | General | Twitter | Not available | New | South Africa |
| [45] | General | | Available | Adopted [82] | Taiwan |
| [85] | General | Facebook /survey | Not available | New | Germany |
| [81] | General | Twitter | Not available | New | Spain |
| [46] | General | Twitter | Available | Adopted [10] | Pakistan |
| [3] | General | Twitter | Available | New | Japan |
| [86] | General | Twitter | Not available | New | Portugal |
| [59] | General | Twitter | Not available | New | India |
| [82] | General | Facebook | Available | New | Taiwan |
| [9] | General | Twitter | Available | New | USA |

down well with many people. Trump is the 6th most followed person on Twitter as of then, with over 87 million followers.[4]

[4]https://en.wikipedia.org/wiki/List_of_most-followed_Twitter_accounts#:∼:text=As%20of%20October%202020%20Barack,account%20with%2087%20million%20followers.

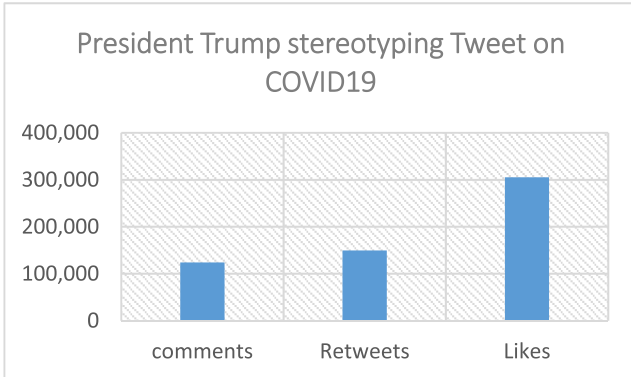


**FIGURE 6.** **President trump tweet analysis.**

The analysis of President Trump tweets is summarized in Figure 6.

The retweets, likes and comments, with over 87 million followers are huge and the impact can be quite devastating to all Chinese across the globe. The centre for disease control (CDC) has cautioned people regarding calling diseases after location, claiming people are been stigmatized.[5] New pandemic or disaster comes with different names which make detection challenging.

## VIII. LIMITATIONS OF THE STUDY

The limitation of this work is that no experiment was conducted with a given dataset. But the work of other researchers was critically appraised. From other researchers works, we were able to synthesis their work and put the conclusion as in the next section.

## IX. CONCLUSION

This article reviewed advances made so far in automatic hate speech detection in social media. Hate speech as a societal problem is an old research area in the arts and humanities, however, it is still a new research area in the computing domain. Therefore, there is a need to constantly update researchers with the advances or progresses made to keep researchers informed. We analysed the approaches from classical ML, Ensemble and deep learning approaches

[5]https://www.nbcnews.com/news/asian-america/trump-tweets-about-coronavirus-using-term-chinese-virus-n1161161

in detecting hate speech in social media. This study found out that there is more research work in hate speech detection using classical ML than ensemble and deep learning techniques. That means researchers can explore more on hate speech detection using ensemble and deep learning methods.

This research also discussed the weaknesses and strengths which can be of help in guiding the researchers' choice of one technique over the other. This article also identified some open challenges in hate speech detection which include: Cultural variations, pandemic or natural disaster, data sparsity, imbalance dataset challenge and dataset availability concern.

The application of ML for automatic HS detection on SM needs to be encouraged and supported. The needs to consider the HS variables based on each country is an issue that needs more researchers' attention. Each country or region may have different variables for HS. For example, marital status and health status are commonly used as HS variable in Nigeria but it has not been addressed by any work in the past.

This research has found out that special characters and numeral symbols mostly used in Nigeria for constructing HS comments have not been addressed by current state-of-the-art. For example, the use of "419" to mean an unwholesome behaviour is commonplace in Nigeria. No research has covered this.

The targeted audience for this research review is mostly newcomers in the domain of hate speech (text) classification in the SM. This review provides all the required steps needed to follow in conducting text classification tasks using ML and some open challenges in the domain.

## CONFLICTS OF INTEREST

There is no conflicts of interest to declare on this study.

## REFERENCES

[1] M. S. Albarrak, M. Elnahass, S. Papagiannidis, and A. Salama, "The effect of Twitter dissemination on cost of equity: A big data approach," *Int. J. Inf. Manage.*, vol. 50, pp. 1–16, Feb. 2020.

[2] C. Cai, H. Xu, J. Wan, B. Zhou, and X. Xie, "An attention-based friend recommendation model in social network," *Comput., Mater. Continua*, vol. 65, no. 3, pp. 2475–2488, 2020.

[3] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate speech on Twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection," *IEEE Access*, vol. 6, pp. 13825–13835, 2018.

[4] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Comput. Surv.*, vol. 51, no. 4, pp. 1–30, Sep. 2018.

[5] A. Guterres, "United nations strategy and plan of action on hate speech," United Nations, New York, NY, USA, Tech. Rep., 2019.

[6] Q. Li *et al.*, *A Survey on Text Classification: From Shallow to Deep Learning*, vol. 37, no. 4. New York, NY, USA: Cornell Univ. Library, 2020.

[7] Q. Al-Maatouk, M. S. Othman, A. Aldraiweesh, U. Alturki, W. M. Al-Rahmi, and A. A. Aljeraiwi, "Task-technology fit and technology acceptance model application to structure and evaluate the adoption of social media in academia," *IEEE Access*, vol. 8, pp. 78427–78440, 2020.

[8] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, pp. 1–68, 2019.

[9] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proc. 11th Int. Conf. Web Soc. Media (ICWSM)*, 2017, pp. 512–515.

[10] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in *Proc. NAACL Student Res. Workshop*, 2016, pp. 88–93.

[11] P. Burnap and M. L. Williams, "Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making," *Policy Internet*, vol. 7, no. 2, pp. 223–242, Jun. 2015.

[12] S. S. Bodrunova, A. Litvinenko, I. Blekanov, and D. Nepiyushchikh, "Constructive aggression? Multiple roles of aggressive content in political discourse on Russian YouTube," *Media Commun.*, vol. 9, no. 1, pp. 181–194, Feb. 2021.

[13] F. Tulkens, "The hate factor in political speech. Where do responsibilities lie?" Polish Ministry Admin. Digitization Council Eur., Warsaw, Poland, Tech. Rep., 2013.

[14] 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, Jan. 2013.

[15] M. A. Al-Garadi, M. R. Hussain, N. Khan, G. Murtaza, H. F. Nweke, I. Ali, G. Mujtaba, H. Chiroma, H. A. Khattak, and A. Gani, "Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges," *IEEE Access*, vol. 7, pp. 70701–70718, 2019.

[16] M. Stegman and M. Loftin, "An essential role for down payment assistance in closing America's racial homeownership and wealth gaps the price of the homeownership gap," Urban Inst., Washington, DC, USA, Tech. Rep., 2021.

[17] R. Alshalan and H. Al-Khalifa, "A deep learning approach for automatic hate speech detection in the Saudi Twittersphere," *Appl. Sci.*, vol. 10, no. 23, pp. 1–16, 2020.

[18] A. Al-Hassan and H. Al-Dossari, "Detection of hate speech in social networks: A survey on multilingual corpus," in *Proc. Comput. Sci. Inf. Technol. (CS IT)*, Feb. 2019, pp. 83–100.

[19] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proc. 5th Int. Workshop Natural Lang. Process. Social Media*, 2017, pp. 1–10.

[20] A. Alrehili, "Automatic hate speech detection on social media: A brief survey," in *Proc. IEEE/ACS 16th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Nov. 2019, pp. 1–6.

[21] A. Rodriguez, C. Argueta, and Y.-L. Chen, "Automatic detection of hate speech on Facebook using sentiment and emotion analysis," in *Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAIIC)*, Feb. 2019, pp. 169–174.

[22] G. Weir, K. Owoeye, A. Oberacker, and H. Alshahrani, "Cloud-based textual analysis as a basis for document classification," in *Proc. Int. Conf. High Perform. Comput. Simul. (HPCS)*, Jul. 2018, pp. 629–633.

[23] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Antisocial behavior in online discussion communities," in *Proc. 9th Int. Conf. Web Soc. Media (ICWSM)*, 2015, pp. 61–70, 2015.

[24] T. Granskogen and J. A. Gulla, "Fake news detection: Network data from social media used to predict fakes," in *Proc. CEUR Workshop*, vol. 2041, no. 1, 2017, pp. 59–66.

[25] L. Tamburino, G. Bravo, Y. Clough, and K. A. Nicholas, "From population to production: 50 years of scientific literature on how to feed the world," *Global Food Secur.*, vol. 24, Mar. 2020, Art. no. 100346.

[26] V. S. Raleigh, "Trends in world population: How will the millenium compare with the past," *Hum. Reprod. Update*, vol. 5, no. 5, pp. 500–505, 1999.

[27] S. Paul, J. I. Joy, S. Sarker, A.-A.-H. Shakib, S. Ahmed, and A. K. Das, "Fake news detection in social media using blockchain," in *Proc. 7th Int. Conf. Smart Comput. Commun. (ICSCC)*, Jun. 2019, pp. 1–5.

[28] World Data Lab. *Population.io by World Data Lab*. Accessed: Jan. 16, 2020. [Online]. Available: https://population.io/?utm_source=google&utm_medium=search&utm_campaign=population&campaignid=1695828135&adgroupid=64502612525&adid=329422103483&gclid=Cj0KCQiAjfvwBRCkARIsAIqSWlN28TwzgVkTSJkTgfnwPfK7fh96cxYU3iglDqWphMuGFdiwTd-04dcaAgofEALw_wcB

[29] G. Liu, C. Wang, K. Peng, H. Huang, Y. Li, and W. Cheng, "SocInf: Membership inference attacks on social media health data with machine learning," *IEEE Trans. Comput. Social Syst.*, vol. 6, no. 5, pp. 907–921, Oct. 2019.

[30] J. van Dijck, "Governing digital societies: Private platforms, public values," *Comput. Law Secur. Rev.*, vol. 36, Apr. 2019, Art. no. 105377.

[31] A. Arango, J. Pérez, and B. Poblete, "Hate speech detection is not as easy as you may think: A closer look at model validation," in *Proc. SIGIR 42nd Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, 2019, pp. 45–53.

[32] C. Ring, "Hate speech IN social media: An exploration of the problem and its proposed solutions," 2013.

[33] S. MacAvaney, H. R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, "Hate speech detection: Challenges and solutions," *PLoS ONE*, vol. 14, no. 8, pp. 1–16, 2019.

[34] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki, "Measuring the reliability of hate speech annotations: The case of the European refugee crisis," 2017.

[35] P. Smith, J. Mahdavi, M. Carvalho, and N. Tippett, "An investigation into cyberbullying, its forms, awareness and impact, and the relationship between age and gender in cyberbullying," Res. Brief, London, U.K., Tech. Rep. RBX03-06, Jul. 2006, pp. 1–69.

[36] M. Yao, C. Chelmis, and D.-S. Zois, "Cyberbullying ends here: Towards robust detection of cyberbullying in social media," in *Proc. World Wide Web Conf. (WWW)*, 2019, pp. 3427–3433.

[37] 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.

[38] K. R. Purba, D. Asirvatham, and R. K. Murugesan, "A study on the methods to identify and classify cyberbullying in social media," in *Proc. 4th Int. Conf. Adv. Comput., Commun. Autom. (ICACCA)*, Oct. 2018, pp. 1–6.

[39] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, "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, Feb. 2014.

[40] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, "Resources and benchmark corpora for hate speech detection: A systematic review," *Lang. Resour. Eval.*, vol. 55, no. 2, pp. 477–523, Jun. 2021.

[41] C. C. Aggarwal and C. X. Zhai, "A survey of text classification algorithms," *Min. Text Data*, vol. 9781461432, pp. 163–222, Feb. 2012.

[42] A. Oma, T. A. El-Hafeez, and T. M. Mahmoud, *Comparative Performance of Machine Learning and Deep Learning Algorithms for Arabic Hate Speech Detection in OSNs*, no. 1153. Cham, Switzerland: Springer, 2020.

[43] W. Dorris, R. Hu, N. Vishwamitra, F. Luo, and M. Costello, "Towards automatic detection and explanation of hate speech and offensive language," in *Proc. 6th Int. Workshop Secur. Privacy Anal.*, Mar. 2020, pp. 23–29.

[44] M. Moh, T. S. Moh, and B. Khieu, "No 'love' lost: Defending hate speech detection models against adversaries," in *Proc. 14th Int. Conf. Ubiquitous Inf. Manag. Commun. (IMCOM)*, Jan. 2020, pp. 1–6.

[45] Z. Mossie and J.-H. Wang, "Vulnerable community identification using hate speech detection on social media," *Inf. Process. Manage.*, vol. 57, no. 3, May 2020, Art. no. 102087.

[46] M. Sajjad, F. Zulifqar, M. U. G. Khan, and M. Azeem, "Hate speech detection using fusion approach," in *Proc. Int. Conf. Appl. Eng. Math. (ICAEM)*, Aug. 2019, pp. 251–255.

[47] D. Suryakumar, A. H. Sung, and Q. Liu, "Determine the critical dimension in data mining (experiments with bioinformatics datasets)," in *Proc. 11th Int. Conf. Intell. Syst. Design Appl.*, Nov. 2011, pp. 481–486.

[48] N. Sharma and K. Saroha, "Study of dimension reduction methodologies in data mining," in *Proc. Int. Conf. Comput., Commun. Autom.*, May 2015, pp. 133–137.

[49] E. N. Sathishkumar, K. Thangavel, and T. Chandrasekhar, "A novel approach for single gene selection using clustering and dimensionality reduction," vol. 4, no. 5, pp. 1540–1545, 2013, *arXiv:1306.2118*. [Online]. Available: https://arxiv.org/abs/1306.2118

[50] L. Nanni, S. Brahnam, C. Salvatore, and I. Castiglioni, "Texture descriptors and voxels for the early diagnosis of Alzheimer's disease," *Artif. Intell. Med.*, vol. 97, pp. 19–26, Jun. 2019.

[51] Z. Zhang and L. Luo, "Hate speech detection: A solved problem? The challenging case of long tail on Twitter," *Semantic Web*, vol. 10, no. 5, pp. 925–945, Sep. 2019.

[52] B. Vidgen and T. Yasseri, "Detecting weak and strong islamophobic hate speech on social media," *J. Inf. Technol. Politics*, vol. 17, no. 1, pp. 66–78, Jan. 2020.

[53] H. Faris, I. Aljarah, M. Habib, and P. Castillo, "Hate speech detection using word embedding and deep learning in the arabic language context," in *Proc. 9th Int. Conf. Pattern Recognit. Appl. Methods*, Jan. 2020, pp. 453–460.

[54] O. Oriola and E. Kotze, "Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets," *IEEE Access*, vol. 8, pp. 21496–21509, 2020.

[55] I. Aljarah, M. Habib, N. Hijazi, H. Faris, R. Qaddoura, B. Hammo, M. Abushariah, and M. Alfawareh, "Intelligent detection of hate speech in arabic social network: A machine learning approach," *J. Inf. Sci.*, May 2020, Art. no. 016555152091765. [Online]. Available: https://journals.sagepub.com/doi/pdf/10.1177/0165551520917651

[56] E. Ombui, L. Muchemi, and P. Wagacha, "Hate speech detection in code-switched text messages," in *Proc. 3rd Int. Symp. Multidisciplinary Stud. Innov. Technol. (ISMSIT)*, Oct. 2019, pp. 1–6.

[57] K. Nugroho, E. Noersasongko, M. Purwanto, A. Z. Fanani, and R. S. Basuki, "Improving random forest method to detect hatespeech and offensive word," in *Proc. Int. Conf. Inf. Commun. Technol. (ICOIACT)*, Jul. 2019, pp. 514–518.

[58] N. Albadi, M. Kurdi, and S. Mishra, "Are they our brothers? Analysis and detection of religious hate speech in the Arabic Twittersphere," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 69–76.

[59] R. Martins, M. Gomes, J. J. Almeida, P. Novais, and P. Henriques, "Hate speech classification in social media using emotional analysis," in *Proc. 7th Brazilian Conf. Intell. Syst. (BRACIS)*, Oct. 2018, pp. 61–66.

[60] P. Burnap and M. L. Williams, "Us and them: Identifying cyber hate on Twitter across multiple protected characteristics," *EPJ Data Sci.*, vol. 5, no. 1, pp. 1–5, Dec. 2016.

[61] A. Géron, *Hands-On Machine Learning With Scikit-Learn and TensorFlow: Concepts. Tools, and Techniques to Build Intelligent Systems*. CA, USA: O'Reilly Media, 2017.

[62] M. Hosni, I. Abnane, A. Idri, J. M. Carrillo de Gea, and J. L. Fernández Alemán, "Reviewing ensemble classification methods in breast cancer," *Comput. Methods Programs Biomed.*, vol. 177, pp. 89–112, Aug. 2019.

[63] P. Montebruno, R. J. Bennett, H. Smith, and C. V. Lieshout, "Machine learning classification of entrepreneurs in British historical census data," *Inf. Process. Manage.*, vol. 57, no. 3, May 2020, Art. no. 102210.

[64] Z. Ding, *Diversified Ensemble Classifiers for Highly Imbalanced Data Learning and their Application in Bioinformatics*. Atlanta, GA, USA: Georgia State Univ., 2011.

[65] A. Suárez-García, M. Díez-Mediavilla, D. Granados-López, D. González-Peña, and C. Alonso-Tristán, "Benchmarking of meteorological indices for sky cloudiness classification," *Sol. Energy*, vol. 195, pp. 499–513, Jan. 2020.

[66] Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, S. Nasrin, and V. K. Asari, "Comprehensive survey on deep learning approaches," 2017.

[67] L. Aristodemou and F. Tietze, "The state-of-the-art on intellectual property analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data," *World Pat. Inf.*, vol. 55, pp. 37–51, Dec. 2018.

[68] Y. Zhang, Q. Wang, Y. Li, and X. Wu, "Sentiment classification based on piecewise pooling convolutional neural network," *Comput., Mater. Continua*, vol. 56, no. 2, pp. 285–297, Jan. 2018.

[69] R. Ong, "Offensive language analysis using deep learning architecture," 2019.

[70] P. M. Sosa, "Twitter sentiment analysis using combined LSTM-CNN models," *Academia.edu*, Jun. 2017.

[71] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proc. 26th Int. Conf. World Wide Web Companion (WWW) Companion*, 2017, pp. 759–760.

[72] G. Kovács, P. Alonso, and R. Saini, "Challenges of hate speech detection in social media," *Social Netw. Comput. Sci.*, vol. 2, no. 2, pp. 1–15, Apr. 2021.

[73] S. Zimmerman, C. Fox, and U. Kruschwitz, "Improving hate speech detection with deep learning ensembles," in *Proc. 11th Int. Conf. Lang. Resour. Eval. (LRED)*, 2019, pp. 2546–2553.

[74] P. Kapil and A. Ekbal, "A deep neural network based multi-task learning approach to hate speech detection," *Knowl.-Based Syst.*, vol. 210, Dec. 2020, Art. no. 106458.

[75] A. Botchkarev, "New typology design of performance metrics to measure errors in machine learning regression algorithms," *Interdiscipl. J. Inf., Knowl. Manage.*, vol. 14, no. 113, pp. 13–21, 2019.

[76] V. Balakrishnan, S. Khan, T. Fernandez, and H. R. Arabnia, "Cyberbullying detection on Twitter using big five and dark triad features," *Personality Individual Differences*, vol. 141, pp. 252–257, Apr. 2019.

[77] H. H. Do, P. Prasad, A. Maag, and A. Alsadoon, "Deep learning for aspect-based sentiment analysis: A comparative review," *Expert Syst. Appl.*, vol. 118, pp. 272–299, Mar. 2019.

[78] C. Bhagat and D. Mane, "Survey on text categorization using sentiment analysis," *Int. J. Sci. Technol. Res.*, vol. 8, no. 8, pp. 1189–1195, 2019.

[79] B. Zhang, S. Zhou, L. Yang, J. Lv, and M. Zhong, "Study on multi-label classification of medical dispute documents," *Comput., Mater. Continua*, vol. 65, no. 3, pp. 1975–1986, 2020.

[80] M. Dholvan, A. K. Bhuvanagiri, and S. M. Bathina, "Offensive text detection using temporal convolutional networks," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 6, pp. 5177–5185, 2020.

[81] J. C. Pereira-Kohatsu, L. Quijano-Sánchez, F. Liberatore, and M. Camacho-Collados, "Detecting and monitoring hate speech in twitter," *Sensors*, vol. 19, no. 21, pp. 1–37, 2019.

[82] Z. Mossie and J.-H. Wang, "Social network hate speech detection for amharic language," in *Proc. Comput. Sci. Inf. Technol.*, Apr. 2018, pp. 41–55.

[83] N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung, "Multilingual and multi-aspect hate speech analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 4667–4676.

[84] K. Dong, T. Guo, X. Fang, Z. Ling, and H. Ye, "Estimating the number of posts in Sina Weibo," *Comput., Mater. Continua*, vol. 58, no. 1, pp. 197–213, 2019.

[85] C. Wilhelm, S. Joeckel, and I. Ziegler, "Reporting hate comments: Investigating the effects of deviance characteristics, neutralization strategies, and users' moral orientation," *Commun. Res.*, vol. 47, no. 6, pp. 921–944, 2019.

[86] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, and M. Shrivastava, "A dataset of hindi-english code-mixed social media text for hate speech detection," in *Proc. 2nd Workshop Comput. Modeling People's Opinions, Personality, Emotions Social Media*, 2018, pp. 36–41.

[87] C. Udanor and C. C. Anyanwu, "Combating the challenges of social media hate speech in a polarized society: A Twitter ego Lexalytics approach," *Data Technol. Appl.*, vol. 53, no. 4, pp. 501–527, Oct. 2019.

[88] M. Luo, K. Wang, Z. Cai, A. Liu, Y. Li, and C. F. Cheang, "Using imbalanced triangle synthetic data for machine learning anomaly detection," *Comput., Mater. Continua*, vol. 58, no. 1, pp. 15–26, 2019.



**NANLIR SALLAU MULLAH** (Member, IEEE) received the B.Tech. degree in computer science from Abubakar Tafawa Balewa University, Bauchi, Nigeria, in 2006, the Post-Graduate Diploma degree in education from Usman Danfodio University, Sokoto, Nigeria, in 2012, and the M.Sc. degree in computer science from Coventry University, U.K., in 2016. He is currently pursuing the Ph.D. degree in computer science with Universiti Sains Malaysia, Penang, Malaysia. From 2007 to 2009, he was a part-time Lecturer with the Plateau State Polytechnic, Barkin-Ladi, Nigeria, and the Federal College of Land Resources Technology, Kuru, Nigeria. Since 2009, he has been an Academic Staff with the Federal College of Education, Pankshin, Nigeria. His research interest includes data science, social computing, text mining, machine learning, big data mining, social media mining, opinion mining, and sentiment analysis.

**WAN MOHD NAZMEE WAN ZAINON** received the Ph.D. degree in computer science from Universiti Sains Malaysia. He is currently a Senior Lecturer with the School of Computer Sciences, Universiti Sains Malaysia. His research interests include visual computing, data mining and software engineering with a focus on software reuse, requirement engineering practices, visual data mining, and big data analytics.