Hate Speech and Abusive Laungage Detection In Twitter and Challenges: Review

Ruba M. Alhejaili
Department of Computer Science and Engineering
College of Computer Science,
Taibah University
Madinah, Saudi Arabia
ruba.alhejaili@gmail.com

Wael M.S. Yafooz
Department of Computer Science and Engineering
College of Computer Science,
Taibah University
Madinah, Saudi Arabia
wyafooz@taibahu.edu.sa

Abdullah A. Alsaeedi
Department of Computer Science and Engineering
College of Computer Science,
Taibah University
Madinah, Saudi Arabia
aasaeedi@taibahu.edu.sa

2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES) | 978-1-6654-8004-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/CISES54857.2022.9844317

***Abstract*—The web' content is increasing every day, especially on social media, where all users can express their opinions freely and without restrictions. Accordingly, many negative activities emerged, such as abusive language, racism, and hate speech. Hate speech or abusive language is a manifestation of negative social media that requires tools to detect it. This work provides a comprehensive view of the concepts of abusive language and hate speech . The methods used for detection will also be presented. The paper also reviews previous studies in this field. Finally, the paper recalls the challenges facing the detection of abusive language or hate speech.**

Keywords—Social Networks, Natural Language, Processing, Classification, Text Mining, Hate Speech.

I. INTRODUCTION

In the twentieth century, the internet and digital technologies emerge, which turned the world into one small village. With such advancements, many applications appeared that made people go about their daily life in a way easier than before. Many of these apps were used by people for ease of use or to reduce the time and effort expended by some traditional methods such as social media [1] and [2] Social media has become one of the easiest ways to follow the latest news around the world, communicate with others, and provide opinions and comments. Social media has become a common ground for people to share their opinions and express their thoughts in a good or bad way.

However, there are some who express their opinions or ideas in a negative way, using some hateful or offensive words. Not only that but some people express their opinion in the form of hate speech against an individual or group. Hate speech, as defined by the authors [3] is an act that detracts from a different individual or group in terms of gender, race, religion, or nationality. Many people believe that expression in all its forms, whether in a good way or not, is protected by freedom of expression. But the use of hate speech or that offensive language may cause psychological, emotional in specific, and social harm in general.

Therefore, detecting hate speech may provide several social benefits to the individual or society. Thus, limiting hate speech may provide individuals with the option to protect themselves from exposure to that harmful content. The rest of paper organize as follows. In Section II, we present the concepts of abusive language and hate speech. In Section III and Section IV, we discuss the machine learning and deep learning algorithms. In Section V and Section IV, we show the literature reviews of detection of hate speech and abusive language using machine learning and deep learning. We present the summary of revised paper in tables in Section VII. In section VIII, we discuss the most important challenges facing the detection of abusive language hate speech. Finally, we conclude this paper in section IX.

II. HATE SPEECH AND ABUSIVE LANGUAGE

Individuals have used social media to express themselves, discuss certain events, and communicate with others. These platforms are not free of offensive or harmful content, so they have become unsafe environments. Lots of harmful content on social media, such as offensive language, extremism, religious bias, sports violence, and hate speech. These individuals believe they are guaranteed freedom of expression. Here arose the problem of balancing freedom of expression and how to limit these abuses and offensive content. Thus, how to control what is published on social media while providing the right to freedom of expression.

Abusive language defined as “an expression of that contains abusive/dirty words or phrases, both oral or text” [4]. To date, there enough international official definition of hate speech. This term is widespread, especially in social media. It refers to the expression of some individuals expressing offensive or hatred and discrimination against some groups different from them.

Hate speech has also been known to be formed through the Internet. When some people use a language that contains speech directed to different individuals and groups to incite hatred, or maybe this language vulgar or obscene or rude comments, so that it results in serious consequences, whether on individuals or societies [5]. Hate speech is any act that may devalue others based on certain characteristics such as sexual orientation, race, ethnicity religion,, gender, and nationality [6]. It also defined hate speech as speech that is directed

towards a person or group, 7 so that the members of the group share specific characteristics that distinguish them from others. [8].

III. MACHINE LEARNING

Supervised Machine Learning is a kind of machine learning algorithm. It is used to categorize items based on predefined labels [9] . There are several literature reviews that have been given attention to using supervised machine learning algorithms in detecting hate speech. In addition, this type is widely used in detecting offensive language or hate speech. This type of model relies on the manual normalization of a dataset. The tagging or labeling process may take a long time and effort, but this may provide a highly efficient classification. Figure 1 shows machine learning process. Below we provide many of the supervised machine learning algorithms.

A. Support Vector Machine (SVM)

Support Vector Machine is an algorithm used to classify texts. It classifies the data in the dataset into classes. Then specifies data points that are close only to the corresponding class. The specified data points are called support vectors. Support vectors are assisted in the classification process. They represent the important points in the dataset. In the SVM algorithm, it tries to find the hyperplane, which characterizes the distance between classes [5].

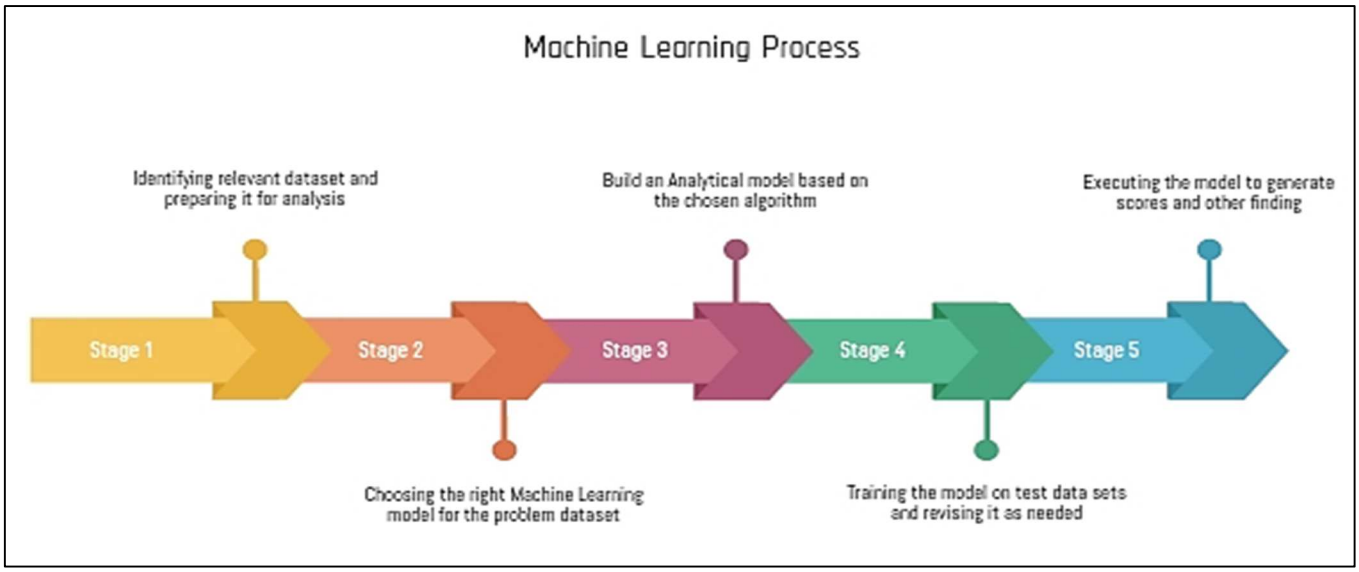

Figure 1: Machine learning process

B. Naïve Bayes (NB)

Naive Bayes is a kind of simple classification model used to classify texts. This algorithm is based on total probability theory and Bayes' theory, which is based on the assumption of conditional independence of traits. With a dataset, the sum of the probabilities can be calculated by the process of calculating values and frequencies. NB calculates the probability of an entry that is relevant to a previously defined class. NB works well for large datasets as it achieves good results [5].

C. K-nearest neighbor (kNN)

kNN is the k-nearest-neighbor to be used in case finding distance is required. After calculating the new instance distance for all instances in the training set it will be possible to predict the class for that new instance [7]. This way the k instances are found which have the smallest distance relative to the new instance whose distance was calculated. Depending on the class designations that are predominant for the closest k-training instances, the same class is assigned to the new instance.

D. Random Forest (RF)

Random forest is a recursive method. Random forests depend on a combination of individual decision trees in the prediction process. Random forests collect and integrate multiple patterns (e.g., trees) within a single random forest. The compilation process is done by creating bootstrapped copies of the original data. Then during each bootstrapped one tree is estimated. One of the most important characteristics of random forests is the low risk of overfitting, with less than one tree. The reason is that the average of multiple trees may provide greater accuracy [10].

E. Decision Tree (DT)

The learning process in the decision tree is by predicting the model that creates the chart or mapping. These diagrams are between observations and conclusions regarding a specific object to its desired value. Therefore, decision trees are among the classification tree models, as they are models that each have specific and fixed values. A decision tree consists of internal nodes, each node representing a feature of an object. As for the terminal nodes, they represent the classifications. Decision trees are used for any type of problem. The most important decision tree method is the classification and regression tree (CART), in which the goal values are continuous. Other decision tree methods include Iteration Dichotomous 3 (ID3), Conditional Decision Trees, and Decision Stump [11].

F. Logistic Regression

Logistic regression is kind of the models that used in classification. It gives probability results according to the values of the input variables. Logistic regression gives a binomial value, either yes or no, and as such, either 0 or 1. In addition, Logistic regression may give a result with multiple limits. Where Logistic regression deals with determining the required variable that represents a specific category [12].

IV. Deep Learning

Deep learning models are part of artificial neural networks. Such models are known as the most important method that used in data mining and classification processes. These models try to learn to accurately identify patterns in texts. The correct choice of any deep learning model depends on how much numbers of hidden layers and the technique of representing the specific features as shown in Figure2. Below we review some literature reviews in this field.

A. Feedforward Neural Network

In a feeding neural network, the information flow is only in one direction. The process of information flow is through the input layer to the output layer (via hidden nodes if there is). The nodes in this type do not constitute any circuits or loopbacks. There is a specific kind of neural network implementation that is characterized by its multilayers as well as values and functions that are computed along the forward path of information. Z represents the weighted sum of the input. y represents the nonlinear activation function f of Z for each layer. W reflects the weights between the two units in adjacent layers, where it is indicated by a set of subscript letters and b represents the unit bias value [13].

B. Recurrent Neural Network (RNN)

In this type of network, the processing units modulate the output. An RNN can have a group of inputs to form a set of outputs [13].

C. Long term short Memory (LSTM)

As in Feedforward Neural Network, the network does not remember the primary inputs for each new entry. To treat this problem Long Short-Term Memory is used. LSTM is a very popular recurring neural network implementation. LSTM holds back states in reverse Feedforward Neural Network. When LSTM is training, it requires providing blocks of memory to store the previous states. Each memory block contains memory cells. The network temporal states are stored inside these cells. Cells control the flow of signals while regulating them through input, forget, and output gates, where the flow of signals is regulated through them. These gates control all cell contents that are stored, written, or read [13].

D. Convolutional neural network (CNN)

Convolutional Neural Network (CNN) is a kind of deep neural networks which widely used in the image processing type, video recognition, medical image classification, and natural language processing. It is a multilayer feed-forward neural network. It is composed of three stages: convolution, nonlinearity, and pooling [15].

- Convolution: a convolution which is a mathematical-based operation is applied in this stage to the input. This operation uses a convolution filter to extract some features and produce feature maps from the input and then to pass these results to the next layer.
- Nonlinearity: This stage aims to try to include some nonlinear properties in the network. Wherein a nonlinear process is used like ReLu.
- Clustering: This stage aims to decrease the dimensions of feature maps by implementing a function such as maximum clustering or average clustering [14].

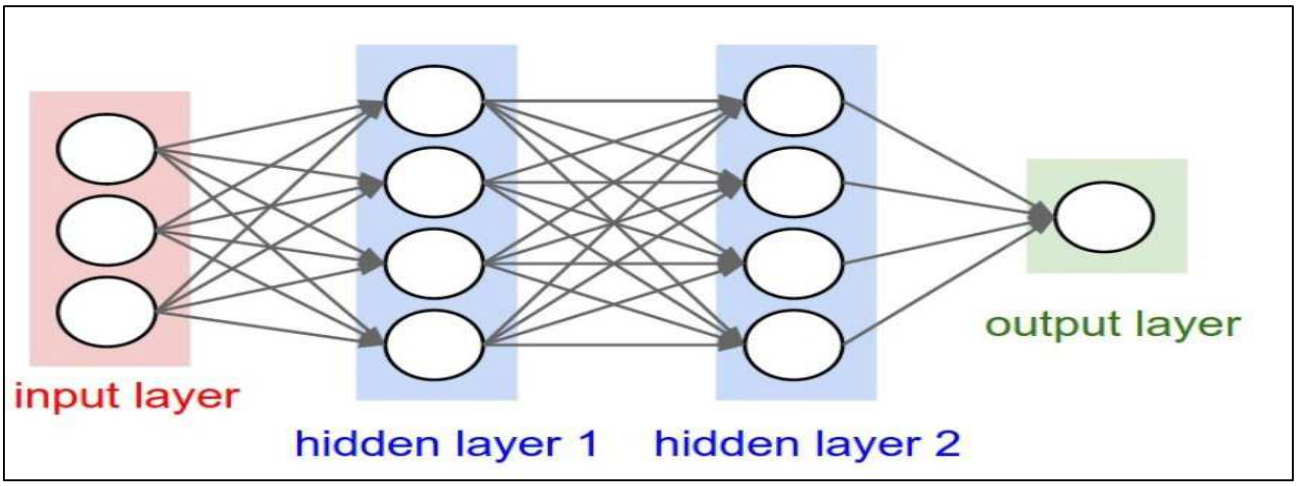

Figure 2: Deep learning model

RNN's sessions and callbacks. Thus, any result of the current layer will be input to the next layer, and so on. The current layer is usually the only layer in the network, so the output of this loop feeds the next loop as input and thus consists of successive loops. This network takes advantage of a memory that stores previous states and uses them to influence current

V. DETECTION USING MACHINE LEARNING

Supervised Machine Learning is a kind of machine learning algorithms. It is used to categories items based on predefined labels. There are several literature reviews has been given attention of using it in detecting hate speech. In addition, this type is widely used in detecting of hate speech or offensive language. This type of model relies on manual normalization of a data set. The tagging or labeling process may take a long time and effort, but this may provide a highly efficient classification. We review some literature reviews that have been based on machine learning models. Some literary studies relied on individual models to classify the required text, as their results proved that the work can achieve good performance. Several previous works have relied on supervised learning techniques that rely on hand naming of texts. The most important of these models is SVM, which usually yields positive results.

For instance, an Arabic dataset has been built on Twitter by [16] on offensive language and hate speech. They used their way to build the data set using manual annotations. They tried a bunch of different dataset models. The result was that SVM gave the best and good results in terms of 88.6% in precision and 79.7% in F1-score.

In addition, in [17] the authors have collected a dataset in the English language to detect hate speech on the social media platforms. They tried to find a distinction between offensive language and the hate speech in general. This data set is divided into three categories: hate, offensive, and Ok (meaning no offensive content). Use the supervised SVM model. To extract the features, they used the three models n-grams, word n-grams, and word skip-grams. The proposed system managed to achieve good performance results on accuracy of 78%.

With the similar results, but with different settings, some authors noticed an increase in hostile behavior in social media, so they tried to detect the abusive language on these platforms [18]. They experimented to detect the offensive Arabic language in comments of the YouTube. They collected and built a dataset, which they believed was the largest dataset composed of YouTube user comments in Arabic at the time. Use the machine learning approach of the SVM model with wordlevel (N-gram) features to enhance the performance of the model. The model performed well with offensive and inoffensive comments classification, with an SVM accuracy of 90.05%. They also concluded that the combination of the SVM with the features of N-gram is useless, and it may have a negative effect.

On the other hand, some literary studies have found that using more than one type of work to classify offensive language and hate speech achieves higher accuracy than if the work used is individual. Among the models most used together are SVM, NB.

To illustrate, in earlier work some authors in [20] used Linear SVM and Naive Bayes machine learning techniques. They used a pre-existing dataset made of tweets, which were divided into three categories: hate speech, offensive language, and normal language. They concluded that Linear SVM rates are low with unbalanced inputs. On the contrary, Naive Bayes does well in classifying texts. Naive Bayes outperformed the Linear SVM with 92% accuracy and 95% recall.

Likewise, the authors followed in [19] built the first Tunisian dataset (T-HSAB) to detect abusive language and hate speech in Tunisian content using the language of machine learning. The dataset was collected and pre-processed from different social platforms, then annotated as normal, abusive, or hate. Using two supervised models NB and SVM to classify toxic content into hate speech (HS) or abusive speech (AS). They made two types of binary classification into normal and abusive by combining hate comments with abusive ones. Also, they used multiple classifications into three categories: normal, abusive, or hate. The two models did well but outperformed the binary classification. Model NB achieved the highest performance in (Acc.), Precision (P.), Recall (R.), F-measure (F1.), and Accuracy. The model results were 93.5% in P., 91.5% in R, 92.3% in F1, and 92.9% in Acc.

Same idea, but with different settings the authors in [21] created a dataset on Twitter and called it L-HSAB. The dataset collected is to try to distinguish offensive language and hate speech in the Arabic language. They categorized the dataset into normal, abusive, and hate. They used two types of models which are SVM and NB. They also use TF and N-GRAM models for word representation. Results were good for both models for precision, recall, accuracy, and F1-score, but NB may outperform SVM where the accuracy rate was up to 90.3%. As observed with the use of the models SVM and NB they achieve high classification accuracy. In the same way, several literary studies followed the same approach but using more types of classification. More than one type of machine learning and ensemble methods were collected, such as DT, RF, LR, and others.

In [22] the authors tried a mission to distinguish hate speech on tweets posted in Portuguese. They used a set of machine learning techniques by defining four models (SVM, MLP, LR, and NB), a pre-equipped dataset made of tweets in Portuguese was used. The dataset was divided into two groups of training and testing, to train the classification models in the two groups using 10-fold cross-validation. Train models based on two categories (hate, no hate) using BoW and N-Gram text representation features. The results of the models used were compared with one of the previous works that performed the same task but using LSTM. They found that the models used outperformed LSTM by about 22% for the SVM model. The SVM model also achieved the best accuracy at 88% and 71% in F1- score. They concluded that classic models may give better and faster results in training.

Similar work, but with different models, the authors in [23] proposed a novel approach to distinguish hate speech in Arabic tweets. The presented approach is based on natural language processing and machine learning methods. They collected and processed the dataset using SVM, NB, DT, and RF. The dataset contained a variety of tweets of a contained racist or related to journalism, sports and terrorism. Features extracted from the dataset were arranged into 15 sub-datasets and varied between words features and emotional features. The features of the extracted words were in three groups of features: BoW, TF, and TF-IDF models. The best model is

RF which reached the highest performance with (TF-IDF) with an accuracy rate of 91%.

Similarly, the authors in [24] used the ensemble method to detect hate speech in the Indonesian language. They choose ensemble classification because they believe it reduces the risk of using a poor model to distinguish hate speech in Tweets. They used the dataset for Indonesian language tweets previously compiled (Alfina et al., 2017). The classification process was done using NB, K-NN, Maximum Entropy, RF, and SVM. The compilation process was divided into two parts, hard voting, and soft voting. Their results showed that the ensemble model has a role in improving data classification performance. The soft vote achieved the best performance score of 79.8% on an unbalanced or unstable dataset and 84.7% on a balanced dataset on an F1 scale.

Moreover, to detect the offensive language in the Indonesian language on social media, the authors in [4] prepared a preliminary study. They also developed a system to detect the offensive Indonesian language. They used a machine learning approach to detect the offensive language in addition to the words' n-gram and char n-gram features. In their paper, they build a new dataset from tweets posted in the Indonesian language. To classify tweets, they used NB, SVM, and Random Forest Decision Tree models (RFDT). They built two experiments on the two-class classification (non-abusive language and offensive language) and the threeclass classification (non-abusive language, abusive but not offensive and offensive language). In their experiment, they reached several results about the performance of the models, where NB was better than SVM and RFDT by 70.06% in the three-class classification and 86.43% in the two-class classification with regard to the F1 score.

In the same way in using a group of models for classification, but according to the content to be classified, in [7] the authors attempted to detect cyberbullying in some social media in the Turkish language. This was the first experience of trying to detect cyberbullying in the Turkish language. Dataset was prepared and collected from Twitter and Instagram comments. To detect cyberbullying, the authors used machine learning models SVM, Decision Tree (C4.5), NBM, and k-NN. The best model in classification performance is Naïve Bayes Multinomial, achieving 84% classification accuracy. Using a group of models gives different results depending on the types of those models or the settings used.

The authors introduced in [25] a mechanism to detect abusive accounts on Twitter through Arabic tweets. That was by categorizing tweets word by word. A dataset of tweets has been compiled and preprocessed containing some swear words in Arabic. They used a set of machine learning algorithms namely NB, SVM, and DT(J48). The results showed that the NB model achieved the best performance with an accuracy of 90%.

VI. DETECTION USING DEEP LEARNING

On the contrary, some literature has turned to deep learning models in classifying offensive language or hate speech. But the results were often not comparable to the accuracy achieved by the supervised learning models. For example, the authors in [26] did the same work [4] to detect the offensive Indonesian language on social media. They were not satisfied with the result that relied on research had achieved, so they used a deep learning approach. They used the same dataset as the original paper to be able to compare the results. To categorize abusive language, they applied long-term memory (LSTM) with word embedding. They find LSTM useful for classification both English and Indonesian when used with word embedding. The model used achieved an 83.68% F1 score, which is about 19.44% higher than the original paper.

Few previous studies have discussed the problem of detecting hate speech during the COVID-19 pandemic. One of those studies was carried out by authors [27] to discuss this problem in the Arab region. They have been using deep learning methods to detect hate speech. Dataset was previously collected in their previous work on Arabic tweets published during the COVID-19 pandemic. In their previous work, they applied CNN's tweet classifying model, which achieved 79% in F1-score and 83% on accuracy. To find out detailed results about the link between hate speech and the Covid-19 pandemic in the Arab region, they analyzed the results from two perspectives, which are the results for the country and for the time. Their results showed that the percentage of hate tweets related to the pandemic was 3.2%, which is much less than the percentage of non-hate tweets (71.4%). Also, their results showed that the highest Arab country in distributing hate tweets through the pandemic was Saudi Arabia.

The same idea, With the increasing use of social networking sites in linguistically different geographic locations, it has become common to hybridize the local language with the English language to facilitate communication with others, such as Henglish. Henglish is a hybrid language between Hindi and English.

Authors in [28] have detect and categorized offensive tweets that were posted in Henglish. In their work, they create a new dataset made of Hindi-English Tweets called the Hindi-English Offensive Dataset (HEOT). After converting the hybrid language to English, they classified the tweets into three categories: non-offensive speech, offensive speech, and hate speech. They used deep learning to classify tweets by training the Convolutional Neural Network Model. Specifically, the triple classification of tweets was achieved using transitional learning on a CNN architecture previously trained on a dataset for further work called the Ternary TransCNN model. The Ternary Trans-CNN model was retrained on a HEOT dataset. By measuring performance in each training, they found Ternary Trans-CNN excelled in the HEOT dataset. The model achieved 83.90% in accuracy and 71% in the F1 Score.

However, some literary studies have used more than one type of classification model. This is in order to improve the accuracy of classification for instance, authors in (Faris et al., 2020) using a deep learning approach to detecting hate speech on Twitter. They collected a dataset of Arabic-language tweets about hate speech. To categorize Tweets, they used a hybrid model, which is a convolutional neural network (CNN) and long short-term memory (LSTM) network. To extract the features, they used the word embedding methods word2vec and Aravec. The classification performed well in

terms of categorizing tweets into hate or normal. Results were better with word2vec, as it achieved 71.688% on the F1 scale and 66.564% on accuracy.

Some studies have used machine learning and deep learning algorithms together. In [29] authors detect hate speech in posted tweets using RF, Complement NB, DT, SVM, and two deep learning methods CNN and RNN. Two types of word embedding methods are also used, FastText and word2vec. They also claim the first Dataset tweets about Sunnis and Shiites (SSTD) have been collected. The results of the deep learning methods were greater than the classic machine learning methods, with the word embedding methods FastText and word2vec. The best accuracy was 79% achieved by CNN with word2vec. The best recall was 69%, which was achieved by CNN with fasttext. The best F1- score was achieved by RNN with Fasttext, which reached 52%.

At the same idea but different settings, on hate speech in the Arabic language, the authors [30] have developed a quick and simple approach to address the problem of distinguishing hate speech in the Arabic language. They compiled and pre-processing the dataset through Arabic. The approach presented adopted several simple features to represent the terminology contained in the Tweets. They compared 15 models that varied between classical and neural models. Their results showed that the neural models exceeded the classical models. The best performing model was the neural learning model (a common structure between CNN and LSTM) that gave a 73% F1-score.

Also, to find and identify people who spread hate, the authors [31] used a deep learning approach to detect hate speech on Twitter. They developed a hybrid technique using a deep neural network that combines convolution, i.e. long-term memory algorithm (LSTM) and convolutional neural network (CNN). For the dataset, they collected it through Twitter, which they preprocessed. Then they trained the model to classify tweets as hate or non-hate. They also created a detailed report on the owners of those tweets that were classified as hate and information about them to track them. These users were identified through the user-id. The report contained the locations of those users, the number of their followers, and their biographical information.

Similarly, the authors [32], using deep learning models, developed a system for classifying hate speech on Twitter. The dataset was used after preprocessing and then they categorized the tweets into one of four predefined categories: racism, sexism, both of them, or neither. Using the Convolutional Neural network model (CNN) that was trained in four models to extract features: Random vectors, word2vec, Character n-grams, and (word2vec + character n-grams). CNN's best performance was with word2vec, which scored 78.3% in F1. Table 1 presents the summary of literature studies that have been discussed.

VII. THE SUMMARY OF REVISED PAPERS

The following table provides a review of the papers reviewed in detecting abusive language and hate speech using machine learning and deep learning techniques. The Table 1 is divided into the part concerned with machine learning methods and the part that is completed by deep learning methods. Each section presents the authors' major findings and what datasets were used in their paper.

VIII. CHALLENGES IN DETECTION HATE SPEECH AND ABUSIVE LANGUAGE

Even with different types of methods for detecting abusive language or hate speech, there are still some challenges facing this field. The most important of these challenges is the difference in the use of words, as some words may be classified as abusive or hateful, but in fact, sometimes they are in the context of speech and may not reflect abusive or hateful content. Also, some speech containing emoji expressions may reflect the understanding of the content of the text. For example, if the text appears as if it does not contain any word that may indicate offensive or hate speech, but it may contain emoji indicating that, here a problem may occur in classifying the content of the text.

Also, the most important of these challenges is the different dialects in the same language, such as the Levantine dialect, the Gulf dialect. While some words can be considered offensive or hateful words in a dialect or in a particular country, they may be classified as a normal word for another dialect or country.

I. CONCLUTION AND FUTURE WORK

Social media is an open field for everyone who wants to write and for everyone who wants to read. Therefore, there is an urgent need to pay attention to social media and what it contains people sharing their opinions and expressing their ideas freely. One of the most important areas that we should pay close attention to is hate speech and abusive language on these programs. In this work, the most important methods and methods that carry out the detection process were presented with a review of some of the previous literary works that discussed the detection of offensive language and hate speech in the means of social communication. Some challenges facing researchers in the detection process were also raised. In the future, a detailed review of detection methods and their appropriate languages will be provided.

Table 1. Summary of the revied papers that presents detection of hate speech and abusive language using machine learning and deep learning.

Authors	Approach	Finding	Dataset
Detection using machine learning			
Mubarak et al., 2020	Machine learning (DT, RF, GNB, AdaBoost, LR and SVM)	SVM had best results.	Dataset of Arabic tweets .
Malmasi & Zampieri, 2017	Machine learning (supervised SVM)	The proposed system managed to achieve good performance results on accuracy of 78%.	English dataset with three categories: hate, offensive, and Ok
Alakrot et al., 2018	Machine learning approach of the SVM model with wordlevel (N-gram) features	The model performed well with offensive and inoffensive comments classification, with an SVM accuracy of 90.05%.They also concluded that the combination of the SVM with the features of N-gram is useless, and it may have a negative effect.	Arabic dataset of YouTube user comments.
Araujo De Souza & Da Costa Abreu, 2020	Linear SVM and Naive Bayes machine learning techniques	Linear SVM rates are low with unbalanced inputs. Naive Bayes does well in classifying texts.	Pre-existing dataset made of tweets, which were allocated into three categories: hate speech, offensive language, and normal language.
Haddad et al., 2019	Supervised models NB and SVM	The two models did well but outperformed the binary classification. Model NB achieved the highest performance in (Acc.), Precision (P.), Recall (R.), F-measure (F1.), and Accuracy	Tunisian dataset (T-HSAB) annotated as normal, abusive, or hate.
Mulki et al., 2019	SVM and NB with TF and N-GRAM	Results were good for both models for precision, recall, accuracy, and F1-score.	Dataset of Twitter and called it L-HSAB. Arabic language with three categories (normal, abusive, and hate).
Silva & Roman, 2020	SVM, MLP, LR, and NB and BoW and N-Gram as a text representation.	The results of the models used were compared with one of the previous works that performed the same task but using LSTM, SVM outperformed by about 22%. Classic models may give better and faster results in training.	Dataset of tweets posted in Portuguese.
Aljarah et al., 2020	SVM, NB, DT, and RF with BoW, TF, and TF-IDF	The best model is RF which accomplished the highest performance with (TF-IDF) with an accuracy rate of 91%.	Dataset of Arabic tweets of a contained racist or related to journalism, sports and terrorism.
Fauzi & Yuniarti, 2018	NB, K-NN, Maximum Entropy, RF, and SVM. The compilation process was divided into two parts, hard voting, and soft voting.	The ensemble model has a role in improving data classification performance.	Dataset of Indonesian tweets.

Ibrohim & Budi, 2018	NB, SVM, and Random Forest Decision Tree models (RFDT).	NB was better than SVM and RFDT by 70.06% in the three-class classification and 86.43% in the two-class classification with regard to the F1 score.	Dataset of Indonesian tweets.
Özel et al., 2017	SVM, Decision Tree (C4.5), NBM, and k-NN.	The best model in classification performance is Naïve Bayes Multinomial, achieving 84% classification accuracy.	Dataset of Turkish language from Twitter and Instagram comments.
Abozinadah et al., 2015	NB, SVM, and DT(J48).	The results indicated that the NB model achieved the best performance with an accuracy of 90%.	Dataset of Arabic tweets.
Detection using deep learning			
Ibrohim et al., 2019	long-term memory (LSTM) with word embedding.	They find LSTM useful for classification both English and Indonesian when used with word embedding. The model used achieved an 83.68% F1 score, which is about 19.44%.	Dataset of Indonesian language on social media.
Alshalan et al., 2020	CNN model.	CNN achieved 79% in F1-score and 83% on accuracy.	Dataset of Arabic tweets.
Mathur et al., 2018	Ternary Trans-CNN.	Ternary Trans-CNN excelled in the HEOT dataset. The model achieved 83.90% in accuracy and 71% in the F1 Score.	Dataset of tweets called the Hindi-English Offensive Dataset (HEOT).
Faris et al., 2020	(CNN) and long short-term memory (LSTM) network.	The classification performed well in terms of categorizing tweets into hate or normal. Results were better with word2vec, as it achieved 71.688% on the F1 scale and 66.564% on accuracy.	Dataset of Arabic-language tweets.
Aref et al., 2020	RF, Complement NB, DT, SVM, and two deep learning methods CNN and RNN. Embedding methods (FastText and word2vec).	The best accuracy was 79% achieved by CNN with word2vec. The best recall was 69%, which was achieved by CNN with fasttext. The best F1- score was achieved by RNN with Fasttext, which reached 52%.	Dataset tweets about Sunnis and Shiites (SSTD).
Abuzayed & Elsayed, 2020	15 models that varied between classical and neural models.	The best performance model was the neural learning model (a common structure between CNN and LSTM) that gave a 73% F1-score	Dataset of Arabic tweets.
Chaudhari et al., 2019	Hybrid technique CNN and LSTM.	These users were identified through the user-id. The report contained the locations of those users, the number of their followers, and their biographical information.	Dataset of English tweets.
Gambäck & Sikdar, 2017	CNN (Random vectors, word2vec, Character n-grams, and (word2vec + character n-grams)).	CNN's best performance was with word2vec, which scored 78.3% in F1.	Dataset of tweets.

REFERENCES

[1] Alsaeedi, A., & Khan, M. Z. (2019). A study on sentiment analysis techniques of Twitter data. International Journal of Advanced Computer Science and Applications, 10(2), 361-374. https://doi.org/10.14569/IJACSA.2019.0100248.

[2] Yafooz, W. M. & Alsaeedi, A. (2021). Sentimental Analysis on Health-Related Information with Improving Model Performance using Machine Learning. Journal of Computer Science, 17(2), 112-122. https://doi.org/10.3844/jcssp.2021.112.122.

[3] Al-Hassan, A., & Al-Dossari, H. (2019). Detection of hate speech in social networks: a survey on multilingual corpus. In 6th International Conference on Computer Science and Information Technology

[4] Ibrohim, M. O., & Budi, I. (2018). A dataset and preliminaries study for abusive language detection in Indonesian social media. Procedia Computer Science, 135, 222-229.

[5] Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S. G., Almerekhi, H., & Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. Human-centric Computing and Information Sciences, 10(1), 1.

[6] Blaya, C. (2019). Cyberhate: A review and content analysis of intervention strategies. Aggression and violent behavior, 45, 163-172.

[7] Özel, S. A., Saraç, E., Akdemir, S., & Aksu, H. (2017, October). Detection of cyberbullying on social media messages in Turkish. In 2017 International Conference on Computer Science and Engineering (UBMK) (pp. 366-370). IEEE.

[8] ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. arXiv preprint arXiv:1804.04257.

[9] Yafooz, W., Emara, A. H. M., & Lahby, M. (2022). Detecting Fake News on COVID-19 Vaccine from YouTube Videos Using Advanced Machine Learning Approaches. In Combating Fake News with Computational Intelligence Techniques (pp. 421-435). Springer, Cham.

[10] Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised machine learning: a brief primer. Behavior Therapy, 51(5), 675-687.

[11] Pandey, D., Niwaria, K., & Chourasia, B. (2019). Machine Learning Algorithms: A Review. Machine Learning, 6(02).

[12] Ray, S. (2019, February). A quick review of machine learning algorithms. In 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon) (pp. 35-39). IEEE.

[13] Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. IEEE Access, 7, 53040-53065.

[14] Fahad, S. A., & Yafooz, W. M. (2017). Review on semantic document clustering. International Journal of Contemporary Computer Research, 1(1), 14-30.

[15] Kim Y (2014) Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

[16] Mubarak, H., Rashed, A., Darwish, K., Samih, Y., & Abdelali, A. (2020). Arabic offensive language on twitter: Analysis and experiments. arXiv preprint arXiv:2004.02192.

[17] Malmasi, S., & Zampieri, M. (2017). Detecting hate speech in social media. arXiv preprint arXiv:1712.06427.

[18] Alakrot, A., Murray, L., & Nikolov, N. S. (2018). Towards accurate detection of offensive language in online communication in arabic. Procedia computer science, 142, 315-320.

[19] Haddad, H., Mulki, H., & Oueslati, A. (2019, October). T-HSAB: a tunisian hate speech and abusive dataset. In International Conference on Arabic Language Processing (pp. 251-263). Springer, Cham.

[20] Da Costa Abreu, M., & Araujo De Souza, G. (2020). Automatic offensive language detection from Twitter data using machine learning and feature selection of metadata.

[21] Mulki, H., Haddad, H., Ali, C. B., & Alshabani, H. (2019, August). L-HSAB: a levantine Twitter dataset for hate speech and abusive language. In Proceedings of the Third Workshop on Abusive Language Online (pp. 111-118).

[22] Silva, A., & Roman, N. (2020, October). Hate Speech Detection in Portuguese with Naïve Bayes, SVM, MLP and Logistic Regression. In Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional (pp. 1-12). SBC.

[23] Aljarah, I., Habib, M., Hijazi, N., Faris, H., Qaddoura, R., Hammo, B., ... & Alfawareh, M. (2020). Intelligent detection of hate speech in Arabic social network: A machine learning approach. Journal of Information Science, 0165551520917651.

[24] Fauzi, M. A., & Yuniarti, A. (2018). Ensemble method for indonesian twitter hate speech detection. Indonesian Journal of Electrical Engineering and Computer Science, 11(1), 294-299.

[25] Abozinadah, E. A., Mbaziira, A. V., & Jones, J. (2015). Detection of abusive accounts with Arabic tweets. Int. J. Knowl. Eng.-IACSIT, 1(2), 113-119.

[26] Ibrohim, M. O., Sazany, E., & Budi, I. (2019, March). Identify abusive and offensive language in indonesian twitter using deep learning approach. In Journal of Physics: Conference Series (Vol. 1196, No. 1, p. 012041). IOP Publishing.

[27] Alshalan, R., Al-Khalifa, H., Alsaeed, D., Al-Baity, H., & Alshalan, S. (2020). Hate Detection in COVID-19 Tweets in the Arab Region using Deep learning and Topic Modeling. Journal of Medical Internet Research. (Preprint).

[28] Mathur, P., Shah, R., Sawhney, R., & Mahata, D. (2018, July). Detecting offensive tweets in hindi-english code-switched language. In Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media (pp. 18-26).

[29] Aref, A., Al Mahmoud, R. H., Taha, K., & Al-Sharif, M. HATE SPEECH DETECTION OF ARABIC SHORTTEXT.

[30] Abuzayed, A., & Elsayed, T. (2020, May). Quick and simple approach for detecting hate speech in arabic tweets. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection (pp. 109-114).

[31] Chaudhari, S., Chaudhari, P., Fegade, P., Kulkarni, A., & Patil, R. (2019, November 11). Hate Speech And Abusive Language Suspect Identification And Report Generation. International Journal of Scientific & Technology Research. https://www.ijstr.org/final-print/nov2019/Hate-Speech-And-Abusive-Language-Suspect-Identification-And-Report-Generation.pdf

[32] Gambäck, B. and Kumar Sikdar, U., 2017. Using Convolutional Neural Networks To Classify Hate-Speech.