Survey on Twitter Sentiment Analysis: Architecture, Classifications, and Challenges

Laith Abualigah, Nada Khaleel Kareem, Mahmoud Omari, Mohamed Abd Elaziz, and Amir H. Gandomi

1 Introduction

Sentiment analysis (SA) is that area for studying knowledge or interpreting people's opinions toward a particular topic [1]. It is also called opinion mining because it interprets the speaker's opinion on a specific topic [2]. It has various names and include different tasks; among those are affect analysis, opinion mining, sentiment mining, subjectivity analysis, and others. Each name has its own job and its diverse tasks, but they all meet to obtain the feelings and opinions of people on a specific topic [1]. In other words, it determines whether the opinion toward a specific topic is negative, positive, or neutral. Hence, sentiments are classified into three categories: negative, positive, and neutral sentiments [2]. Positive sentiments are the good terms about the topic in consideration. When the positive impressions are high, it concludes good feelings. Negative sentiments, on the other hand, are the bad terms about the topic in consideration. For example, many business owners use Twitter to track and monitor people's opinions about their products and services. When positive feedbacks about a product are high, then the expected purchase rate would be high. On the other hand, when negative impressions are high, it is rejected from the

L. Abualigah (✉)
Faculty of Computer Sciences and Informatics, Amman Arab University, Amman, Jordan

School of Computer Sciences, Universiti Sains Malaysia, Pulau Pinang, Malaysia

N. K. Kareem · M. Omari
Faculty of Computer Sciences and Informatics, Amman Arab University, Amman, Jordan

M. A. Elaziz
Department of Mathematics, Faculty of Science, Zagazig University, Zagazig, Egypt

A. H. Gandomi
Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW, Australia

V. Kadyan et al. (eds.), *Deep Learning Approaches for Spoken and Natural Language Processing*, Signals and Communication Technology,
https://doi.org/10.1007/978-3-030-79778-2_1

preference list, and no purchases are expected for the product. Finally, neutral sentiments are neither good nor bad terms about the topic. Hence, it is neither favor nor neglected.

SA is a natural language processing (NLP) task that extracts natural data in which it focuses on obtaining feelings. In general, it focuses on inferring from the behavior of the speaker or person regarding a particular topic [3]. The field of sentiment analysis is multidisciplinary, as it deals with individuals' feelings, opinions, emotions, and attitude toward products, services, topics, issues, individual, and anything else subject to opinion. Subjects of sentiment analysis include various areas like computer languages, NLP, machine learning (ML), and technical intelligence as well as information retrieval. It includes a set of computational and natural languages based on techniques that can be used to extract data from a specific text diverse to subjective opinions or feelings [4]. The field of SA is a subdomain of ML. With manual training, the problem can be solved to analyze feelings to the required level as there is no automated integrated system for analyzing feelings without requiring manual intervention [5].

Different levels can be applied to sentiment analysis, including the *document level* that gives one polarity to the entire document, the *sentence level* that gives polarity to each sentence, and the *entity/aspect level* that is based on analyzing each word that has feelings. Previously, sentiment analysis was limited to knowing the polarity of opinions. It was of no use in making or taking decisions, as no reasons were known about why the sentiments have changed. Hence, there is a need to build systems to explain these differences in public sentiments. Several studies of multiple sentimental techniques and various algorithms are used to analyze feelings [2, 3]. The main resource for sentiment analysis is web data, which is huge and the largest store of unstructured and structured information.

Nowadays, social networks have become widely used. Facebook, Twitter, LinkedIn, and YouTube in addition to other sites are very popular. Twitter is the most used platform in social media, which is a microblog that permits its users to post their feelings and opinions. The number of Twitter registrants in 2017 reached nearly 696 million, and the amount of tweets per day reached approximately 58 million tweets. Such microblogging has become an essential and important source of great value to people's opinions and feelings. It includes various topics and different directions, including political and economic in addition to religious and social as well as sports and other trends. Such important data can be used efficiently in studying market conditions, social studies, disease surveillance, and other common topics. Twitter users are not only, regular users but also leaders heads of state, firm's executives, and celebrities [1, 6]. For example, if you want to know the people's opinion about the former US President Barack Obama, you can refer to social network sites such as Twitter. The Twitter platform includes millions of opinions about what Obama has done during his presidency. You would find positive, negative, and neutral opinion tweets. Accuracy can be obtained in the answer to whether the people believe that he fulfilled his duties or not by extracting some accurate words from the tweets that indicate the opinion on this matter [2].

Obviously, it is necessary to collect and analyze the data represented by text posts on various matters in order to reach the feelings expressed in the tweets. Most of the data available in networks is irregular, accounting for approximately 80% of all data in the world. It is difficult to obtain accurate information, make a judgment, or analyze this data. Sentiment analysis is of great importance in extracting or mining opinions, as it helps to reveal people's feelings or opinions from social media, which is used for sharing opinions and ideas between people linked on the global web [1, 6]. The importance of this study stems from the importance of SA to all people's endeavor [2]. This study will describe the process of SA and its dependency on various levels of text analysis, namely, sentence level, document level, and phrase level. We will also look at the architecture of the sentiment analysis process, which includes four stages: data collection, preprocessing, feature extraction, and training and classification. We will also discuss more broadly the most important algorithms used in this field. The challenges that face sentiment analysis will be described. Finally, we will remark some gaps that assist to expand the scope of progression in this research area.

This chapter is organized as follows. Section 2 describes the different levels of sentiment analysis. Section 3 describes the architecture used for performing a comprehensive and detailed sentiment analysis. The various algorithms used in this field are discussed in Sect. 4. In Sect. 5, we will discuss some of the challenges facing SA. Section 6 concludes the study and gives further research direction.

2 Levels of Sentiment Analysis

Various studies on sentiment analysis have been done at three main levels [1–4]:

(a) *Document level analysis:* In this type [7], the entire document is viewed, and the goal is to analyze the overall feel of the document as the whole document is viewed as one subject. For example, when reviewing a specific product, the task will be to determine whether opinions are negative, positive, or neutral for that product. Sometimes, this level does not give correct results, as is the case in forums or blogs where one product is compared to another for the same characteristics. In addition, it is possible that the document contains sentences that are not related to the topic and should not be included in the process of analyzing feelings.

(b) *Phrase level analysis:* This level is limited to analyzing a complete sentence and decides whether it has positive, negative, or neutral feelings. A sentence may have two or more words. It is close to subjective classification; it carries accurate data from sentences that include the opinions and objective aspect. It may give inaccurate results if the sentence includes a negative sentence.

(c) *Entity/aspect level analysis:* It is also called the feature level. Instead of analyzing a document or sentence, this level is based on analyzing each word that has feelings. This level provides an accurate analysis of each aspect (feature level). It

analyzes the word to obtain different opinions. The word may be an adjective or an adverb and may be a noun. It depends on a concept; the opinion can be an attitude, a word, or a point of view.

3 System Architecture

Sentiment analysis is based on words that people use to express their opinion in a specific matter (negative, positive, or neutral). To reveal the direction of view, we first identify the meaningful words in the tweets and then discover their direction whether that word reflects positive feelings, negative feelings, or neutral feelings [8]. The system comprises four key stages: data collection, data processing, and feature extraction [1], as explained in Fig. 1. Data are obtained or collected from Twitter and preprocessed, which includes filtering to filter unique Twitter features and extracting side-based features to identify explicit and implicit aspects [9]. As an input, the system collects the tweet identifiers and the N-Gram model for the purpose of learning a class [10]. The tweets are normalized and converted to mono grams (four grams, three grams, two grams, and monograms) [11]. After preparing training and testing data, the different classifiers are applied to analyze the performance of the workbooks [6] and thus obtain the outputs. In the following, we discuss each stage in detail.

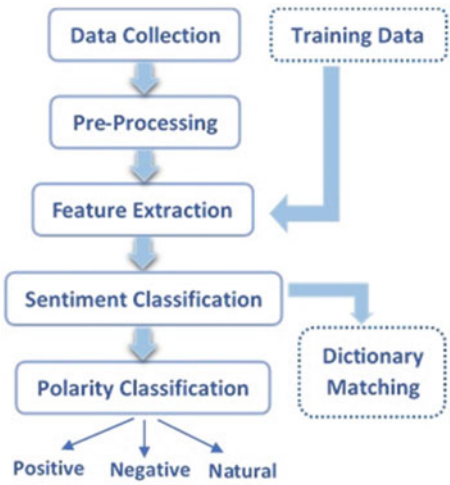

Fig. 1 Sentiment analysis architecture

3.1 Data Collection

It is important to collect data for a specific topic from relevant sites (as in microblogging) using queries [6]. This means that if we need to analyze feelings of opinion in the health field, for example, it is necessary to rely on sites specific to this field [12]. Blogs are sites that enable users to post messages, pictures, links to other sites, or videos. Messages posted on blogs are short, unlike traditional blogging. Currently, a number of platforms are available for blogs, including Twitter, LinkedIn, Google, Foursquare, and Tumblr [13]. Here, we will talk about Twitter as a source of data.

3.1.1 Twitter

Twitter, one of the most popular microblogging sites, is a form of platform and microblogging that permits its users to post messages denoted as tweets. Tweets contain numerous unique features [8, 14]. Twitter began in 2006 and has since attracted a large number of users. Ease of accessing and downloading publications and the amount of data it contains. Twitter was considered one of the largest datasets [13] that could be adopted in sentiment analysis. The adoption of Twitter data in the analysis of feelings is to classify the tweets into different feelings categories accurately [1]. Twitter data was adopted in various fields, as it was used as a source to monitor real-world outcomes or forecast information, including the analysis of extreme events such as Syria 2013, expected box office revenue for films, earthquakes in Japan, and others [12]. Twitter is characterized by some specific features that are listed below:

- *Tweet:* Tweet is the message that was posted on Twitter. The maximum message limit is approximately 140 characters. Tweet content (tweet topics) can differ from a personal opinion on a specific topic or news and may be in the form of links, news, and photos or may include videos.
- *Writing technology:* Because messages are short, many use abbreviations in comments, and some use symbols that give a lot of meaning. This is in addition to spelling errors, incorrect spelling in many tweets, and use of colloquial language.
- *Availability:* The number of people tweeting in the public domain on Twitter is large compared to other platforms such as Facebook (Facebook has many privacy settings). This made data widely available, as it is easy to collect tweets for training.
- *User/username*: When registering in the system, a name is chosen, and the name can be a pseudonym. This name is used in the system to post tweets.
- *Mention:* The Mention is when the tweet is being referred to another user, to share the topic with that user. The tweet uses the "@" symbol before the username that is indicated (@username).

- *Comments:* Comments often create conversation, which is the result of answering a comment, and other people are referred to.
- *Follower:* Followers are the ones who follow the user and his activities. Follow-up is the way to communicate with other Twitter users. Where the user receives an update from the followers and also sends his updates to the followers.
- *Retweet:* When a tweet is posted, another user can re-publish that tweet using retweet. It is considered a strength for disseminating information. It can be seen that the tweet was reposted with the abbreviation RT followed by a username (RTusername).
- *Hashtag:* This feature is used to classify the tweet and its relevance to a specific topic. The symbol # is followed by the name of the topic (#topic). The Hashtag is used by the tweet, thus access to all tweets using the same hashtag. Classification Hashtag are often popular topics.
- *Privacy:* This feature determines whether the tweet will be visible to everyone or only for followers. All of these characteristics that are mentioned are problems; on the other hand, these problems need to be processed, which we will address in the second stage [1, 8, 13, 14].

3.2 Preprocessing

Twitter tweets were adopted because they contain many opinions, which are presented in various ways. It has been categorized into positive and negative as well as neutral, which makes data analysis not difficult [15]. The representation of tweets is often in vague and informal ways [11]. And because of the diversity of language usage in tweets, it is possible that there are language or spelling errors. Tweets can include some symbols, abbreviations, usernames, links, and others that are not related to the classification process [9]. Therefore, processing techniques are used to obtain relevant content, while the rest of the comments away from the topic are ignored. This stage is very important in the classification process [15]. Hence, data quality has a significant impact on the results. So to enhance the analysis, the preliminary data are processed [15]. Among the most important processing steps that are implemented are as follows:

- *Tokenization:* After the tweets are compiled with identifiers available in the data sets, each tweet is broken down into a set of individual words. For each tweet, there will be a list of its own individual words [16–18].
- *Removal of non-English tweets:* The nature of the Twitter allows the use of more than 60 languages. The focus will be on the English language. We will remove non-English words and tweets.
- *Replace emoticons in many microblogging posts:* Many Twitter users use emoticons and shortcuts for tweets. Each of these symbols has strong connotations and is an indicator of feelings, as they are a concise way of identifying feelings. It will therefore have a vital part in determining the feelings of the tweet. It will be

Table 1 Some emoticons and their meaning

Emoticon	Meaning	Sentiment Class
:-D	Laughing	Positive
:-)	smile	Positive
o:)-	innocent	Positive
8-)	cool	Positive
:$	Happy blush	Positive
:(	defeated	Negative
:'(	Crying	Negative
:o	shocked	Negative
>(	Grumpy	Negative
(@)	Angry red	Negative
X|	Dead	Negative

easy to distinguish the polarity of messages, whether positive or negative. This is done using the emoticon dictionary, where the symbol is replaced with corresponding emotions (Table 1).

- *Removal of links/URL:* Due to the limited length of the text for tweets, users use URLs. These URLs do not carry any meaningful indications, as a word, within the tweet itself. But it does provide a large content of emotion that the user tries to express in a concise manner. However, it remains very difficult to reach the content of the URLs; therefore, the URL will be deleted.
- *Removal of target mentions:* Most of the time the user mentions another user in the tweet. It can be distinguished by the "@" symbol. It is placed in front of the username you want to refer to (@John). This part of the tweet (@John) is not important in the analysis because it does not have any moral significance. So it will be removed.
- *Removal of punctuations from hashtags:* The hashtags are important, providing a summary of what the tweet means. So to get the information, you must delete both punctuation and the symbol indicating it to retain only the important information from it.
- Removal of numbers: Sometimes, numbers are used in tweets. The numbers have no value when measuring feelings. Therefore, the numbers are removed from the tweet content.
- Handle sequences of repeated characters: Spelling correction is of great importance in analyzing feelings for tweet content. Often users express their opinions abnormally and loudly, without focusing on the correct texture and spelling. Tweeters use words like "cooooooool" or "wooooooow." In order to get the correct expression for such words, we replace the repeated letter more than three times with three letters to be "coool" and "wooow." We substitute three letters to distinguish between emphasized usage of the word and regular usage of the word; for example, the word "cool" gives another meaning. WordNet is used to ensure that unnecessary characters are removed.

- *Removal of stop words:* The tweets include many words that have no meaning, called stop words. Stop words do not contain any information about feelings and therefore are useless. When making tokenization, these words should be removed. Examples of stop words are a, an, the, and other words.
- Handle negative mentions: Negativity has an important and significant role in determining the tweet. The words "no," "not," "never," and others or the words that end with it should be replaced by a word referring to negation.
- *Uppercase identification:* It is common to use capital letters to express strong emotions. Such a type is called an e-shouting. It is a good indicator to get the message polarity easily. This step mines this feature before taking out casing [10, 11, 19, 20].

Note that we will need resources to process Twitter data including an emoticon dictionary and an acronym dictionary [21]. The dictionary of emoticon is for emoticons, which includes a number of the most used emoticons. As for the dictionary of acronym that is compiled from various sources and which includes translations of a large number of abbreviations, we will mention them in another section of this work.

3.3 Feature Extraction

The identification and selection of features are very significant for the classification of texts. We try to understand which features are the most significant for the classification process. Text feature extraction is the process of obtaining a list of words from previously processed data and then converting that list into a set of features that can be used by the classifier. A variety of methods for defining and extracting features from textual data are important for classification; some of these features are as follows [4, 19, 22, 23]:

- *Part-of-speech tagging (PoS)*: Important signals of opinion to find an adjective or a descriptive word for each sentence. The natural language processing technique uses pointers to the parts of speech. PoS indicator is an undertaking for the labeling of all words in a sentence to a PoS tag. These words relate to communal categories of English grammar including adjectives, verbs, names, and prepositions, as well as conjugation, pronouns, and interference. Because parts of speech define expressions of feelings and semantic relationships between expressions, they are used to filter features that indicate the direction of feelings.
- *N-gram model*: A set of texts related to a subject is analyzed. The texts represent tweets. Each word, or symbol, is extracted from tweets as a series of words, which is represented as N. At the end, a dictionary of words or symbols is formed. That sequence of symbols or words can be a letter, word, or byte and can be continuous symbols. Depending on the number of grams, the words are defined, which means 1-gram, which is also named a unigram, is composed of one symbol; 2-gram or bigram consists of two symbols, and trigram consists of three symbols. This

makes the analysis process capable of revealing the correlation between those words and the prominence of the phrase itself. For instance, the text "Microsoft is launching a new product" is composed of the resulting 2-gram word features: "Microsoft is," "is launching," "launching a," "a new", and "new product." The tweet is represented in N-gram, as in the previous example. These features are N-gram words or solo words with their repeatedly counts. The tweet features will be a series of 1s and 0s. 1 represents that the Tweet contains N-gram, while 0 indicates its absence.

- *Unsupervised feature weighting methods*: Weighting techniques can be classified into two main classes: unsupervised and supervised techniques. The supervised technique utilizes previous data from the learning document to formulate a group of pre-produced classifies. This differ from the supervised feature weighting technique. Among the techniques utilized under unsupervised weighting are the TF-IDF (term frequency-inverse document frequency) and binary term frequency (TF).
- *Dimensionality reduction utilizing principal component analysis (PCA)*: It is considered a common method for extracting features. It has been applied in many fields and on wide and varied groups in various fields of biological and social sciences to the field of financing. As this approach is based on setting data points. To determine those points that are more important in that space, two criteria are applied: the proportion of variance and the Kaiser rule.
- *Word2vec model*: Word2Vec is used to create word embeddings. The models formed by using word2vec are little denotation two-layer neural networks. Once learned, they propagate semantic cases of words. The model takes a vast frame of text as a feed-in. It then constructs a vector scope that is usually of hundreds of dimensions. Each special word in the frame is allocated with symmetric vector in the scope. The words with common cases are placed in near closeness in vector scope. Word2vec uses one of the two constructions: continuous skip gram, which considered the current word is to forecast the neighboring window of case words. In this construction, the nearby case words are treat more constructions than words with outlying case. Or continuous bag of words (CBOW), the series of case words does not affect the prognosis as it is founded on bag of words sample.
- *Bag of words*: It is considered one of the simplest and most common methods of extracting features as this method is flexible. It is used to extract features from text data in various ways. Each word has a group of similar words; it is collected inside a bag called a word bag. WordPad is a display of textual data, which determines the frequency of words in a document. The word bag includes a dictionary of well-known words and the frequent presence of those words. The complexity of this model of feature extraction lies in the degree to which those words are present as well as how vocabulary is designed for those words. Despite the ease and flexibility of this model, word repetition is a problem that cannot be overlooked. The data with the highest frequency will be in control of the rest of the bag data. Higher frequency data may not be important, or model information may not be available. This problem is the main reason for ignoring related words.

- *TF-IDF*: In this way, we will be able to display unique words that carry the necessary information for a single document.
- *Opinion words and phrases:* They are phrases generally utilized to indicate opinions that are composed of bad or good, love or hate. This means some words indicate opinions devoid of utilizing opinion phrases.
- *Negations:* The existence of negative words might turn the opinion directing like not bad is equivalent to good.

3.4 Classification

After preprocessing and feature extraction level, we move to classification level. We pass the features into a classifier [10]. Many techniques were built for text classification. In this level, the general classes of techniques will be discussed, as well as their utilization specific to classification tasks. We understand that the discussed classes of techniques largely exist in other domains like categorical data or quantitative [24]. Sentiment classification strategies are classified into ML, lexicon-based approach, and hybrid technique as in [14]. In an ML approach, the popular ML algorithms are used in addition to the semantic indications of these algorithms, including naive Bayes classifier, SVM, decision tree, and others. We can classify machine learning into supervised learning and unsupervised methods. As for the lexicon-based approach, it is solely reliant on the dictionary of feelings, which is a collection of expressions of feelings previously collected. It is divided into two: the dictionary-based method and the corpus-based method. There is the hybrid method that combines the previous two approaches and through which we may get better results. Several optimization techniques can be used to optimize the classification process or feature selection processes [25, 26]. We will look at a number of widely used algorithms in Sect. 4 [14].

4 Sentiment Classification Techniques

In Fig. 2, we illustrate the classification technique. The general classes of techniques and their utilization for classification tasks will be further discussed.

4.1 Machine Learning Approach

An ML-based SA system was progressed in many earlier works to elicit public views in related topics. This system was capable of classifying tweets to various sentiment classes [27]. Text Classification Problem Definition: there is a list of training documents where each one is classified to a class. The classification model is suitable

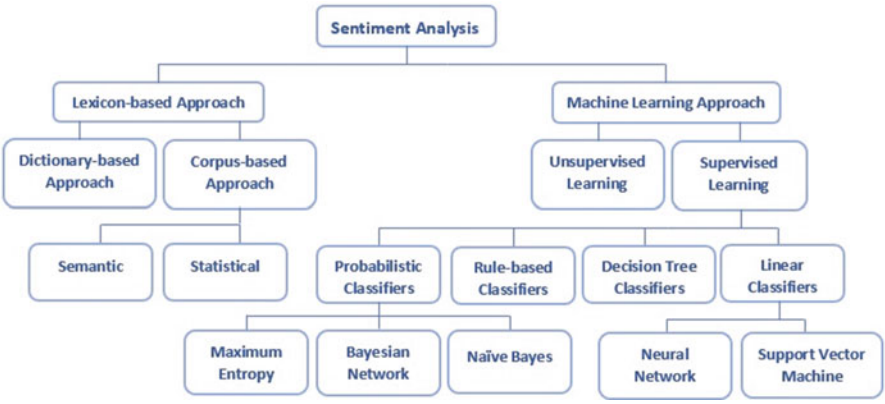

Fig. 2 Tree diagram of sentiment classification techniques

to the features in the inferior record to labeled according to the class label. Next for a given case of an undiscovered class, the model is utilized for the prediction a class label. The severe classification problem is in a situation when solely a single label was chosen to an occurrence. The machine learning technology uses many famous algorithms in machine learning and employs semantic functions [14].

4.1.1 Classification Based on Supervised Learning

Supervised learning is a type of ML [7, 28]. A supervised learning strategy relies on data called training data. Previously classified objects are entered into the device as training data. The device learns from that data. Then it will predict the unclassified data. There are many works under supervision; we will mention some of them [14, 29].

Naive Bayes Classifier

Naive Bayes is a simple, easy yet powerful rule suite [14]. Accordingly, it is used in both the training and classification stage [15]. It is a probabilistic classifier that uses Bayes' theorem to calculate the prospect of a tweet that belongs to a particular category like negative, positive, or neutral [20]. It can train the pattern of test on a group of documents that had been classified. The basic mechanism of the naive Bayes classifier is done by calculating the repetition of words concerning feelings in the message. Tweets are categorized and recorded according to the number of matches for emotional words, as the contents compare the word list to categorize documents into their correct category. The heaviness of the nodes is modified according to the significance of the tweets, and a more exact result can be generated for the rated feelings [2]. Preprocessed data is provided with the extracted feature as

an introduction to training the classifier by utilizing naive Bayes. As soon as the learning is done, in the classification process, it offers polarity of feelings. For instance, for review comment "I am Sad," it provides negative polarity as a result [15]. Naive Bayes classifiers are computational fast while making selections. Naive Bayes classifier is well used in real problems along with spam detection, SA, etc. It demands low processing memory and less execution time. The naive Bayes-type model estimates the backward potential classification based on the word distribution in the report [14].

Support Vector Machine (SVM)

Support vector machine (SVM) is a procedure for categorization of both nonlinear and linear data. SVM is generally used for text classification. If the data is linearly distinct, SVM studies the data, defines the selection limits, and utilizes the kernels for calculation that are executed in input space with the support vector machine.

Searches for the linear optimum split up hyperplane (the linear kernel) that is regarded as the vector, which splits up document vector in one class from the vectors in other class. Furthermore, all data regarded as a vector is categorized in a specific class. Hence, the work is to identify an edge between two categories that is distant from all documents [2, 15, 30]. It uses possibility space of linear functions in a loudly dimensional feature space, depending on kernel substitution. It will be learned with a training algorithm that executes a learning bias derived from statistical training theory. We can construct loudly nonlinear classification method by using SVMs [20]. This is done when the data is linearly close; the SVM utilizes nonlinear chart to turn data to a greater dimension. The problem is then fixed by locating a linear hyperplane [30]. SVM has the ability to replace the teaching styles dynamically each time there may be a new brand pattern throughout category [14].

Neural Network Classifiers

Neural networks (NN) are a specific set of algorithms, which have transformed machine learning [14]. Neural networks are stimulating with the aid of biotic neural networks [14]. The basic element in an NN is a unit or neuron. Hence, each unit takes a certain input i denoted by the vector X_i. Every neuron is also related with a group of weights A, which are utilized to calculate a function f of the inputs. A common function that is frequently utilized in the NN is the linear function stated as follows: $p_i = A * X_i$.

The question here is: how an NN may be utilized if all the categories cannot be carefully divided using linear separator? The categories might not be detached with the utilization of a single linear separator. The utilization of multilayers of neurons can be adopted to create a nonlinear classification limit. The purpose of such multilayers is to create multiple linear limitations that can be utilized to sacrificial bounded parts that belong to a specific class. In a network like this, the outputs of the

neurons in the previous layers are fed into the neurons in the next layers. The learning process of this kind of networks is rather sophisticated because the errors require to be back-propagated over varied layers. However, the general monitoring for text has been that linear categories mostly offer similar results to nonlinear data, and the refinements of nonlinear categorization methods are rather small [24]. Neural networks are criterion feature processes; that is why they can be used in almost any device training problem regarding studying a sophisticated mapping from the entry to the output area. Neural network-based planning has executed brilliant refinements in an expansion of herbal language processing involvement [14].

Random Forest (Decision Tree)

Decision forest provides a more accurate classification than a single decision tree [30] due to the fact that it consists of more than one decision tree [20]. The basis for decision tree is the hierarchical decomposition of data space, where the data is divided repeatedly until we find that each leaf node contains the minimum number of decisions [24]. Here, we will find that each tree will assign a specific category to each entry. The layer with the highest turn will be selected. Error rate depends on the strength of each tree separately from the forest, as well as the relationship between the trees in the forest itself. This means that reducing the error rate is contingent on the strength and independence of each tree in the forest [20]. To illustrate the work of the forest, we will take an illustrative example: X is the main forest, consisting of Y number of sub-trees. Each sub-tree is called X_i. The X_i is made up of branches with the same number of rows of the main X, which are the sample x with the substitution of X. By taking those samples with the substitution, this means that some of the traits in the original sample X may not be included in X_i, while it can be repeated in the others. Then the compiler builds a decision tree for each X_i. And at the end, we will have a forest with Y trees. To categorize an anonymous group, M, each tree brings back its own row prediction as a single vote. Therefore, the last decision of the M class is signed on which the most votes [30].

Rule-Based Classifier

Rules-based methods are based on entity recognition. In general, the nature of rule-based methods works as follows: At first, a group of rules is manually denoted or automatically trained [31]. The rules are adopted in the design of data space. The rule consists of two aspects. The left side called "pattern" represents the basic condition for the set of basic features. The pattern determines the regular expression based on features of tokens. The pattern matches a series of tokens, at which time the specified action is launched. The right side called "action" represents the category designation corresponding to the relevant feature. As for the action, it is possible to name a series of tokens as an entity, specify the beginning or ending designation for the entity, or specify a number of entities simultaneously [24, 31]. For instance, to point out any

series of symbols "Mr. Y" where Y is a capital letter as an individual entity, the subsequent rule can be specified: *(Token = "orthography" orthography type = FirstCap)→the person's name.*

The left side of the rule is a logical condition that can be expressed in DNF (disjunctive normal form). Nevertheless, in many circumstances, the condition on the left side is much plain and signifies a series of terms, which has to be existing in the document in order for the condition to be approved [24]. Generally utilized features to act tokens contain the token oneself, the part-of-speech (POS) identify the token, the orthography kind of the token, and if the token is inside some already defined gazetteer [31]. The lack of expression is hardly utilized, since such kind of rules is not probable but very factual for sparse text data, whereby many statements in the lexicon will normally not exist by default (sparseness property). The basic idea of making a method is to produce a group of rules, where all points are covered in an area of at least one base. An amount of criteria can be utilized to produce rules from learning data. The most common conditions used to create rules are trust and support [24].

Bayesian Classifiers

Bayesian networks (BNs), or belief networks (Bayes networks), are also called generative workbooks. It is one of the types of models with vector graphics. This means that the graphic links are represented by an arrow that indicates a certain direction. Attempting to construct a probability categorization depends on modeling keyword features in different categories. The idea of classifying texts is formed via the rear possibility of documents that belong to distinct categories, and this is based on the existence of a word in those documents. English language weights have an effect on the existence of words through many input groups. As the representations of the subject model are attractive, they get information that is absent in other methods. For instance, for documents that must be summarized, there will be an unambiguous representation as different documents to form that group. In usual ways, multiple document entries will be represented as a long text without differentiating the document limitations. Hence, for Bayesian classifiers, the rear possibilities are weighted through the cost of the category where the prediction is conducted [24, 32].

4.1.2 Classification Based on Unsupervised Learning

Unsupervised learning is a kind of machine learning. This type of machine learning relies heavily on speculative density in statistics. There is no goal in unsupervised learning, but it provides space for an evaluation of the model. The area can check a given model by relying on data such as the input values that are passed to the model [14, 29].

4.2 Lexicon-Based Approach

We mentioned in the previous sections the possibility of using the lexicon of feelings, which is one of the most important resources for most of the algorithms of sentiment analysis. Lexicon includes many words of opinion used in many classification tasks. Here, we will briefly mention some methods for creating lexicon of opinion words. Opinion words, polar words, or emotional words are utilized to express chosen situations if the opinion words are positive or unwanted if they are negative. Examples of positive opinion words are beautiful, good, etc. For negative opinion words, examples are bad, poor, etc. Words of opinion will not come at all times in the form of individual words; they can come in the form of opinion phrases or terms. For example, it costs a person an arm and a leg. Jointly, they are named the lexicon of opinion. Where it is important in the analysis of feelings. Words of opinion can be separated into two parts: the first is called the basic type, where all previous instances represent this type, and the second is called the comparative type. The comparative type is based on the principle of comparison and preference in opinion, for example, better, worse, and more. The preferential and comparative forms are used for the basic characteristics or conditions, for example, good and bad. Here, we will focus on the basic type. To compile a list of words of opinion, there are three main methods: the manual approach, the corpus-based approach, and the dictionary-based approach. The manual approach consumes a lot of time, so it is usually not used; in some cases, it is combined with automated methods. Here, we will address the two automated approaches [14, 33, 34]:

- Dictionary-Based Approach

  To obtain polarity of sentiments, a lexicon of opinion is used. The polarity number of negative and positive words shown for each tweet is calculated. This method determines the highest number of polarities. If polarity is equal to both positive and negative, then polarity is considered neutral to that tweet. The purpose of using this method is to facilitate the access to sentiment words with their directions. However, it failed to define the directions that adopt the context formula for sentiment words [6]. Words of opinion are gained manually to create a group of opinion words. This group can be expanded with the help of searching within WordNet to add new opinion words or a list of their synonyms or oppositions. Then the examination is done manually to remove and correct errors. The downside of this type is its weakness in identifying words of opinion in context-specific directives [14, 34].
- Corpus-Based Approach

  In this model, after extracting opinion words from tweets, their direction is determined. Words of opinion will be a mixture of verbs and adjectives in addition to circumstances. Conditions and verbs are not adopted, and to calculate their direction, the dictionary-based method is used. As for adjectives, an adjective is a word used to describe and qualify an object. It is field dependent, where the corpus-based method will be utilized to obtain the semantic direction of the adjectives [30, 34].

5 Challenges Involved in Sentiment Analysis

There are some challenges that should be faced in sentiment analysis. Some of them are listed as follows [14]:

- *Language problem*: The English language is used to analyze sentiment well due to the availability of resources in English that facilitate the analysis process. Resources are lexicons and dictionaries. There are many researchers interested in analyzing sentiment in other languages, including Chinese, Arabic, German, and others. This makes it challenging for researchers to create resources in other languages, i.e., glossary and dictionaries.
- *Natural language processing (NLP)*: The utilization of NLP requires further improvements in the sentiment analysis process, as it has become a magnet for researchers. Natural language processing offers better mining results and also provides a good awareness of the language. Mining for opinion is context based or field based. Opinion mining needs to pay a lot of attention to it because field-based mining provides a good result than context-based mining. Domain-based mining is complicated or more difficult to develop.
- *Fake opinion*: Fake opinion, or fake review, refers to false reviews that mislead consumers by presenting opinions that are not real, negative or positive, in order to reduce the condition of the object. Such SPAM makes sentimentality ineffective in many applications.

6 Conclusion and Future Work

The data that is used to examine sentiment is social networks. Twitter is the most important of them. It is analyzed in different perspectives to express sentiment and opinions. We explain the concept of sentiment and the structure of data processing to extract opinion at various levels of sentiment analysis. The polarity of sentiment was also categorized as negative, positive, and neutral. In addition to subcategories, which are very negative and very positive. Many approaches have been used in the field of SA, including ML and lexicon. The dictionary of abbreviations and emoticons is also used in sentiment analysis. Here, we presented an overview of the most important algorithms that had good results. Naive Bayes and SVM algorithms are commonly utilized to identify the problem of classification of sentiment. SA is a broad field, opens wide means for research fields and various issues. So the field of investment has become large in the SA. Where it is adopted at the political and economic level and many areas. However, there are some challenges facing the SA, one of the most vital of which is limited data resources for non-English languages, as explained in Sect. 5.

Sentiment analysis is a broad and important area. Many institutions in various fields are adopting sentiment analysis in their work. Millions of tweets every day bring up various topics. This diversity in the content of the tweets needs to be

analyzed according to the direction of the field or topic that was raised. There are many methods and algorithms used in the analysis and approved to reach an opinion that is closer to the truth or accuracy in the results. All studies presented dealt with textual sentiment or textual data. The tweets can contain opinions that are presented in the form of pictures or videos or can be in the form of a link to communicate the sentiment or opinions more accurately and clearly. So we recommend that there be studies to analyze the sentiment of images or videos, or possible links, which are an important element in communicating the sentiment directly because of the important feelings that it holds for the opinion.

References

1. R. Wagh, P. Punde, Survey on Sentiment Analysis Using Twitter Dataset, in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, (IEEE, 2018), pp. 208–211
2. B.S. Dattu, D.V. Gore, A survey on sentiment analysis on twitter data using different techniques. Int. J. Comp. Sci. Inform. Technol. **6**(6), 5358–5362 (2015)
3. H. Hajipoor et al., A survey on twitter sentiment analysis. Proceedings of the First International Conference on Web Research (ICWR), Tehran, Iran, 15–16 (2015)
4. G. Beigi, X. Hu, R. Maciejewski, H. Liu, An Overview of Sentiment Analysis in Social Media and Its Applications in Disaster Relief, in *Sentiment Analysis and Ontology Engineering*, (Springer, Boston, MA, 2016), pp. 313–340
5. R. Varghese, M. Jayasree, A survey on sentiment analysis and opinion mining. Int. J. Res. Eng. Technol. **2**(11), 312–317 (2013)
6. A.P. Jain, V.D. Katkar, Sentiments Analysis of Twitter Data Using Data Mining, in *2015 International Conference on Information Processing (ICIP)*, (IEEE, 2015), pp. 807–810
7. L. Abualigah, A.H. Gandomi, M.A. Elaziz, H.A. Hamad, M. Omari, M. Alshinwan, A.M. Khasawneh, Advances in meta-heuristic optimization algorithms in big data text clustering. Electronics **10**(2), 101 (2021)
8. A. Kumar, T.M. Sebastian, Sentiment analysis on twitter. Int. J. Comp. Sci. Issues (IJCSI) **9**(4), 372 (2012)
9. N. Zainuddin, A. Selamat, R. Ibrahim, Hybrid sentiment classification on twitter aspect-based sentiment analysis. Appl. Intell. **48**(5), 1218–1232 (2018)
10. A. Dalmia, M. Gupta, V. Varma, IIIT-H at SemEval 2015: Twitter sentiment analysis–the good, the bad and the neutral! Proc. 9th Int. Worksh. Seman. Eval. (SemEval) **2015**, 520–526 (2015)
11. R. Pandarachalil, S. Sendhilkumar, G.S. Mahalakshmi, Twitter sentiment analysis for large-scale data: an unsupervised approach. Cogn. Comput. **7**(2), 254–262 (2015)
12. V. Carchiolo, A. Longheu, M. Malgeri, Using Twitter Data and Sentiment Analysis to Study Diseases Dynamics, in *International Conference on Information Technology in Bio-and Medical Informatics*, (Springer, Cham, 2015)
13. A. Giachanou, F. Crestani, Like it or not: a survey of twitter sentiment analysis methods. ACM Comput. Sur. (CSUR) **49**(2), 1–41 (2016)
14. C. Bhagat, D. Mane, Survey on text categorization using sentiment analysis. Int. J. Sci. Technol. Res. **8**(8), 1189–1195 (2019)
15. G. Gautam, D. Yadav, Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis, in *2014 Seventh International Conference on Contemporary Computing (IC3)*, (IEEE, 2014)
16. L.M.Q. Abualigah, E.S. Hanandeh, Applying genetic algorithms to information retrieval using vector space model. Int. J. Comp. Sci. Eng. Appl. **5**(1), 19 (2015)

17. L.M. Abualigah, A.T. Khader, Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering. J. Supercomput. **73**(11), 4773–4795 (2017)
18. L. Abualigah, M. Qasim, *Feature Selection and Enhanced Krill Herd Algorithm for Text Document Clustering* (Springer, Berlin, 2019)
19. V.S. Pagolu et al., Sentiment Analysis of Twitter Data for Predicting Stock Market Movements, in *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, (IEEE, 2016)
20. B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N. Prasath, A.S. Perera, Opinion Mining and Sentiment Analysis on a Twitter Data Stream, in *International Conference on Advances in ICT for Emerging Regions (ICTer2012)*, (IEEE, 2012), pp. 182–188
21. A. Agarwal, B. Xie, I. Vovsha, O. Rambow, R.J. Passonneau, Sentiment analysis of twitter data. Proc. Worksh. Lang. Soc. Media (LSM) **2011**, 30–38 (2011)
22. A.G. Shirbhate, S.N. Deshmukh, Feature extraction for sentiment classification on twitter data. Int. J. Sci. Res. (IJSR) **5**(2), 2183–2189 (2016)
23. R.N. Waykole, A. Thakare, A review of feature extraction methods for text classification. IJAERD **5**(04), 351–354 (2018)
24. C. C. Aggarwal, C. X. Zhai (eds.), *Mining Text Data* (Springer Science & Business Media, New York, 2012)
25. L. Abualigah, A. Diabat, S. Mirjalili, M. Abd Elaziz, A.H. Gandomi, The arithmetic optimization algorithm. Comput. Methods Appl. Mech. Eng. **376**, 113609 (2021)
26. L. Abualigah, A. Diabat, Advances in Sine Cosine Algorithm: A comprehensive survey. Artif. Intell. Rev. **54**, 1–42
27. J. Du et al., Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with twitter data. BMC Med. Inform. Decis. Mak. **17**(2), 69 (2017)
28. L. Abualigah, A.H. Gandomi, M.A. Elaziz, A.G. Hussien, A.M. Khasawneh, M. Alshinwan, E.H. Houssein, Nature-inspired optimization algorithms for text document clustering—A comprehensive analysis. Algorithms **13**(12), 345 (2020)
29. T. Upadhya, S. Raj, S. Pathak, Machine learning techniques for code optimization. Mach. Learn. **6**(07) (2019)
30. X. Fang, J. Zhan, Sentiment analysis using product review data. J. Big Data **2**(1), 1–14 (2015)
31. J. Jiang, Information Extraction from Text, in *Mining Text Data*, (Springer, Boston, MA, 2012), pp. 11–41
32. A. Nenkova, K. McKeown, A Survey of Text Summarization Techniques, in *Mining Text Data*, (Springer, Boston, MA, 2012), pp. 43–76
33. R. Feldman, Techniques and applications for sentiment analysis. Commun. ACM **56**(4), 82–89 (2013)
34. B. Liu, Sentiment analysis and subjectivity. Handb. Nat. Lang. Process. **2**(2010), 627–666 (2010)