A Study on Sentiment Analysis Techniques of Twitter Data

Abdullah Alsaeedi[1], Mohammad Zubair Khan[2]
Department of Computer Science,
College of Computer Science and Engineering
Taibah University
Madinah, KSA

***Abstract*—The entire world is transforming quickly under the present innovations. The Internet has become a basic requirement for everybody with the Web being utilized in every field. With the rapid increase in social network applications, people are using these platforms to voice them their opinions with regard to daily issues. Gathering and analyzing peoples' reactions toward buying a product, public services, and so on are vital. Sentiment analysis (or opinion mining) is a common dialogue preparing task that aims to discover the sentiments behind opinions in texts on varying subjects. In recent years, researchers in the field of sentiment analysis have been concerned with analyzing opinions on different topics such as movies, commercial products, and daily societal issues. Twitter is an enormously popular microblog on which clients may voice their opinions. Opinion investigation of Twitter data is a field that has been given much attention over the last decade and involves dissecting "tweets" (comments) and the content of these expressions. As such, this paper explores the various sentiment analysis applied to Twitter data and their outcomes.**

***Keywords*—Twitter; sentiment; Web data; text mining; SVM; Bayesian algorithm; hybrid; ensembles**

I. Introduction

Sentiment analysis is also known as "opinion mining" or "emotion Artificial Intelligence" and alludes to the utilization of natural language processing (NLP), text mining, computational linguistics, and bio measurements to methodically recognize, extricate, evaluate, and examine emotional states and subjective information. Sentiment analysis is generally concerned with the voice in client materials; for example, surveys and reviews on the Web and web-based social networks.

As a rule, sentiment analysis attempts to determine the disposition of a speaker, essayist, or other subjects in terms of theme via extreme emotional or passionate responses to an archive, communication, or occasion. The disposition might be a judgment or assessment, full of emotion (in other words, the passionate condition of the creator or speaker) or an expectation of enthusiastic responses (in other words, the impact intended by the creator or buyer). Vast numbers of client surveys or recommendations on all topics are available on the Web these days and audits may contain surveys on items such as on clients or fault-findings of films, and so on. Surveys are expanding rapidly, on the basis that individuals like to provide their views on the Web. Large quantities of surveys are accessible for solitary items which make it problematic for clients as they must peruse each one in order to make a choice. Subsequently, mining this information, distinguishing client assessments and organizing them is a vital undertaking. Sentiment mining is a task that takes advantage of NLP and information extraction (IE) approaches to analyze an extensive number of archives in order to gather the sentiments of comments posed by different authors [1, 2]. This process incorporates various strategies, including computational etymology and information retrieval (IR) [2]. The basic idea of sentiment investigation is to detect the polarity of text documents or short sentences and classify them on this premise. Sentiment polarity is categorized as "positive", "negative" or "impartial" (neutral). It is important to highlight the fact that sentiment mining can be performed on three levels as follows [3]:

- Document-level sentiment classification: At this level, a document can be classified entirely as "positive", "negative", or "neutral".
- Sentence-level sentiment classification: At this level, each sentence is classified as "positive", "negative" or unbiased.
- Aspect and feature level sentiment classification: At this level, sentences/documents can be categorized as "positive", "negative" or "non-partisan" in light of certain aspects of sentences/archives and commonly known as "perspective-level assessment grouping".

The main objective of this paper is to study the existing sentiment analysis methods of Twitter data and provide theoretical comparisons of the state-of-art approaches. The paper is organized as follows: the first two subsequent sections comment on the definitions, motivations, and classification techniques used in sentiment analysis. A number of document-level sentiment analysis approaches and sentence-level sentiment analysis approaches are also expressed. Various sentiment-analysis approaches used for Twitter are described including supervised, unsupervised, lexicon, and hybrid approached. Finally, discussions and comparisons of the latter are highlighted.

II. Definition and Motivation

Sentiment analysis is a strategy for checking assessments of people or groups; for example, a portion of a brand's followers or an individual customer in correspondence with a customer supports representative. With regard to a scoring mechanism,

sentiment analysis monitors discussions and assesses dialogue and voice affectations to evaluate moods and feelings, especially those associated with a business, product or service, or theme.

Sentiment analysis is a means of assessing written or spoken languages to decide whether articulation is positive, negative or neutral and to what degree. The current analysis tools in the market are able to deal with tremendous volumes of customer criticism reliably and precisely. In conjunction with contents investigation, sentiment analysis discovers customers' opinions on various topics, including the purchase of items, provision of services, or presentation of promotions.

Immense quantities of client-created web-based social networking communications are being persistently delivered in the forms of surveys, online journals, comments, discourses, pictures, and recordings. These correspondences offer significant opportunities to obtain and comprehend the points of view of clients on themes such as intrigue and provide data equipped for clarifying and anticipating business and social news, such as product offers [4], stock returns [5], and the results of political decisions [6]. Integral to these examinations is the assessment of the notions communicated between clients in their content interchanges.

"Notion examination" is a dynamic area of research designed to enhance computerized understanding of feelings communicated in content, with increases in implementation prompting more powerful utilization of the inferred data. Among the different web-based social networking platforms, Twitter has incited particularly far-reaching client appropriation and rapid development in terms of correspondence volume.

Twitter is a small-scale blogging stage where clients generate 'tweets' that are communicated to their devotees or to another client. At 2016, Twitter has more than 313 million dynamic clients inside a given month, including 100 million clients daily [7]. Client origins are widespread, with 77% situated outside of the US, producing more than 500 million tweets every day [8]. The Twitter site positioned twelfth universally for activity in 2017 [9] and reacted to more than 15 billion API calls every day [10]. Twitter content likewise shows up in more than one million outsider sites [8]. In accordance with this enormous development, Twitter has of late been the subject of much scrutiny, as Tweets frequently express client's sentiment on controversial issues. In the social media context, sentiment analysis and mining opinions are highly challenging tasks, and this is due to the enormous information generated by humans and machines [11].

III. Importance and Background

Opinions are fundamental to every single human action since they are key influencers of our practices. At whatever point we have to settle on a choice, we need to know others' thoughts. In reality, organizations and associations dependably need to discover users' popular sentiments about their items and services. Clients use different types of online platforms for social engagement including web-based social networking sites; for example, Facebook and Twitter. Through these web-based social networks, buyer engagement happens progressively. This kind of connection offers a remarkable open door for advertising knowledge. Individuals of every nationality, sexual orientation, race and class utilize the web to share encounters and impressions about virtually every feature of their lives. Other than composing messages, blogging or leaving remarks on corporate sites, a great many individuals utilize informal organization destinations to log opinions, express feelings and uncover insights about their everyday lives. Individuals compose correspondence on nearly anything, including films, brands, or social exercises. These logs circulate throughout online groups and are virtual gatherings where shoppers illuminate and impact others. To the advertiser, these logs provide profound snippets of insight into purchasers' behavioral inclinations and present a continuous opportunity to find out about client emotions and recognitions, as they happen without interruption or incitement. Be that as it may, recent explosions in client-produced content on social sites are introducing unique difficulties in capturing, examining and translating printed content since information is scattered, confused, and divided [12].

Opinion investigation is a method of information mining that can overcome these difficulties by methodically separating and dissecting web-based information without causing delays. With conclusion examination, advertisers are able to discover shoppers' emotions and states of mind continuously, in spite of the difficulties of information structure and volume. The enthusiasm in this study for utilizing sentiment analysis as an instrument for promoting research instrument is twofold.

Sentiment analysis critically encourages organizations to determine customers' likes and dislikes about products and company image. In addition, it plays a vital role in analyzing data of industries and organizations to aid them in making business decisions.

IV. Classification Techniques

In the machine learning field, classification methods have been developed, which use different strategies to classify unlabeled data. Classifiers could possibly require training data. Examples of machine learning classifiers are Naive Bayes, Maximum Entropy and Support Vector Machine [14] [15, 16]. These are categorized as supervised-machine learning methods as these require training data. It is important to mention that training a classifier effectively will make future predictions easier.

A. Naïve Bayes

This is a classification method that relies on Bayes' Theorem with strong (naive) independence assumptions between the features. A Naive Bayes classifier expects that the closeness of a specific feature (element) in a class is disconnected to the closeness of some other elements. For instance, an organic fruit might be considered to be an apple if its color is red, its shape is round and it measures approximately three inches in breadth. Regardless of whether these features are dependent upon one another or upon the presence of other features, a Naïve Bayes classifier would consider these properties independent due to the likelihood that this natural fruit is an apple. Alongside effortlessness, the Naive Bayes is known to out-perform even exceedingly

modern order strategies. The Bayes hypothesis is a method of computing for distinguishing likelihood P(a|b) from P(a), P(b) and P(b|a) as follows:

$$p\left(\frac{a}{b}\right) = \left[p\left(\frac{b}{a}\right) * p(a)\right]/p(b) \quad (1)$$

Where $p\left(\frac{a}{b}\right)$ is the posterior probability of class a given predictor b and $p\left(\frac{b}{a}\right)$ is the likelihood that is the probability of predictor b given class a. The prior probability of class a is denoted as $p(a)$, and the prior probability of predictor p is denoted as $p(b)$.

The Naive Bayes is widely used in the task of classifying texts into multiple classes and was recently utilized for sentiment analysis classification.

B. Maximum Entropy

The Maximum Entropy (MaxEnt) classifier estimates the conditional distribution of a class marked a given a record b utilizing a type of exponential family with one weight for every constraint. The model with maximum entropy is the one in the parametric family $P_{MaxEnt}\left(\frac{a}{b}\right)$ that maximizes the likelihood. Numerical methods such as iterative scaling and quasi-Newton optimization are usually employed to solve the optimization problem. The model is represented by the following:

$$P_{MaxEnt}\left(\frac{a}{b}\right) = \frac{exp[\sum_i \alpha_i f_i(a,b)]}{\sum_a exp[\sum_i \alpha_i f_i(a,b)]} \quad (2)$$

Where a is the class, b is the predictor. The weight of vector is denoted as α_i

C. Support Vector Machine

The support vector machine (SVM) is known to perform well in sentiment analysis [13]. SVM investigates information, characterizes choice limits and uses the components for the calculation, which are performed in the input space [18]. The vital information is presented in two arrangements of vectors, each of size m. At this point, each datum (expressed as a vector) is ordered into a class. Next, the machine identifies the boundary between the two classes that is far from any place in the training samples [19]. The separate characterizes the classification edge, expanding the edge lessens ambivalent choices. As demonstrated in [20], the SVM has been proven to perform more effectively than the Naïve Bayes classifier in various text classification problems.

V. Document-Level Sentiment Analysis Approaches

Sharma *et al.* [2] proposed an unsupervised document-based sentiment analysis system able to determine the sentiment orientation of text documents based on their polarities. This system [2] categorizes documents as positive and negative [2, 3, 19] and extracts sentiment words from document collections, classifying them according to their polarities. Fig. 1 shows a case of document-based opinion mining. The unsupervised dictionary-based strategy is utilized as a part of this system, which additionally takes care of negation. WordNet is a lexicon adopted to define opinion vocabularies, their equivalent words, and antonyms [2]. In this particular study, movie reviews were collected to utilize as input so as to detect the polarity sentiment of documents. The system classified each of them as positive, negative and impartial and generated summary outputs, presenting the total number of positive, negative and nonpartisan documents. Thus, the summary report produced by the system helped decision makers. With this system, the sentiment polarity of any document is decided based on the majority of opinion vocabularies that appear in documents.

Chunxu Wu [21] proposed a method for synthesizing the semantic orientations of context-dependent opinions that cannot be determined using WordNet. The proposed method is utilized to decide the sentiment of opinions by utilizing semantic closeness measures. This approach relies on such measures to determine the orientation of reviews when there is insufficient relevant information. The experiment conducted by Chunxu Wu [21] demonstrated that the proposed procedure was extremely effective.

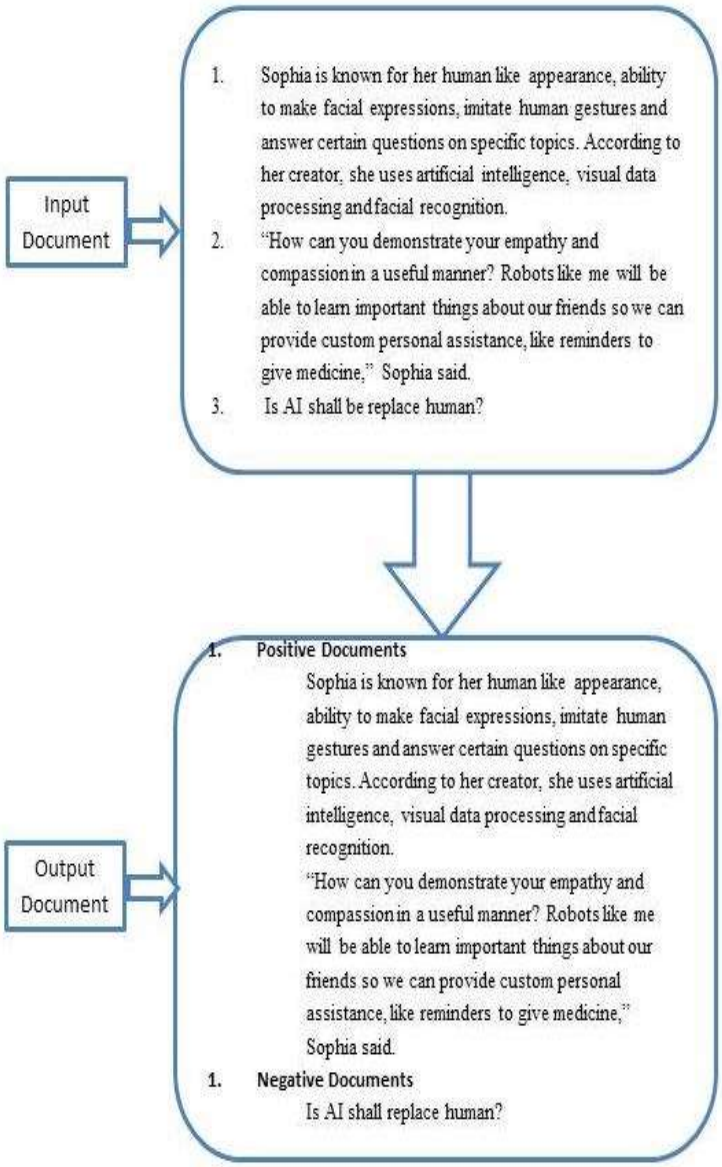

Fig 1. Example of Document-based Opinion Mining.

Taboada *et al.* [22] utilized a lexicon-based technique to detect and classify documents based on their sentiments. To achieve this appropriately, positive and negative word lexicons were utilized. In addition, the semantic orientation calculator (SO-CAL) was proposed, which relies on intensifiers and negation. This SO-CAL approach attained 76.37% accuracy on movie reviews datasets.

Harb *et al.* [18] proposed a document-level sentiment extraction approach concentrating on three stages. In the first stage, a dataset consists of documents containing opinions which have been automatically extracted from the Internet. Secondly, positive and negative adjective sets are extracted from this learning dataset. In the third stage, new document test sets are classified based on adjective lists collected in the second stage. Numerous experiments were conducted on real data and the approach proposed by Harb *et al.* [18] accomplished an F1 score of 0.717 for identifying positive documents and an F1 score of 0.622 for recognizing negative records.

Zagibalov *et al.* [23] addressed the issue of sentiment classification of reviews about products written in Chinese. Their approach relied on unsupervised classification able to teach itself by increasing the vocabulary seed. It initially included a single word (good) that was tagged as positive. The initial seed was iteratively retrained for sentiment classification. The opinion density criterion was then utilized to compute the ratio of sentiments for a document. The experiments showed that the trained classifier attained an F-score of 87% for sentiment polarity detection after 20 iterations.

Tripathy *et al.* [24] attempted to classify reviews according to their polarity using supervised learning algorithms such as the Naïve Bayes, the SVM, random forest, and linear discriminant analysis. To achieve this, the proposed approach included four steps. First, the preprocessing step was carried out to remove stop words, numeric and special characters. Second, text reviews were converted into a numeric matrix. Third, the generated vectors were used as inputs for four different classifiers. The results were subsequently obtained by classification of two datasets. After that, various metrics, such as precision, recall, f-measure, and classification accuracy, were computed to assess the performance of the proposed approach. For the polarity and IMDb datasets, the random forest classifier outperformed other classifiers.

Saleh *et al.* [25] applied the SVM to three different datasets in order to classify document reviews. Several n-grams schemes were employed to evaluate the impact of the SVM in classifying documents. The researchers utilized three weighting approaches to generate feature vectors: namely, Term Frequency Inverse Document Frequency (TFIDF), Binary Occurrence (BO) and Term Occurrence (TO). Numerous experiments were then conducted to measure the possible combinations of various n-grams and weighting approaches. For the Taboada dataset, the best accuracy result was obtained using a combination of the SVM with the TFIDF and trigram. For the Pang corpus, the best results were obtained using the BO and trigram. As regards the SINAI corpus, Saleh *et al.* [25] showed that the SVM classifier achieved the highest accuracy score when combined with the TFIDF and bigram.

VI. Sentence-Level Sentiment Analysis Approaches

This analysis focuses on classifying sentences into categories according to whether these sentences are positive, negative, or neutral. Twitter sentiment analysis is considered an example of sentence-level sentiment analysis. The next section explores Twitter sentiment analysis approaches. Machine learning approaches utilize classification methods to classify text into various categories. There are mainly two types of machine learning strategies: supervised learning and ensemble.

There are four basic Twitter sentiment analysis approaches including supervised machine learning-based, ensemble methods, lexicon-based, and hybrid. These four approaches are described are as follows:

A. Twitter Sentiment Analysis using Supervised Machine Learning Approaches

It depends on labelled datasets that are given to machine learning models during the training process. These marked datasets are utilized to train these models to obtained significant outputs. In machine learning systems, two datasets are required: training set and test set. Machine learning approaches such as classifiers can be utilized to detect the sentiment of Twitter. The performance of Twitter sentiment classifiers is principally relying upon the number of training data and the features sets are extractors. Twitter sentiment analysis strategies that rely on machine-learning methods are more popular, especially SVM and NB classifiers. Fig. 2 illustrates the procedure of supervised machine learning approaches for Twitter sentiment analysis.

The Twitter sentiment analysis process consists of three steps. First, the classifier is trained using datasets comprising positive, negative, and unbiased tweets. Examples of tweets are shown below:

- The following tweets are examples of positive tweets:

1) PM@narendramodi and the President of Ghana, Nana Akufo-Addo had a wonderful meeting. Their talks included discussions on energy, climate change and trade ties.

2) Billy D. Williams @Msdebramaye For the children, they mark, and the children, they know The place where the sidewalk ends.

3) @abdullah “Staying positive is all in your head”. #PositiveTweets

- Unbiased tweets

1) (@Nisha38871234): "#WorldBloodDonorDay Blood Donation is the best donation in the world. Save a life!!"Good night #Twitter and #TheLegionoftheFallen. 5:45am cimes awfully early!

2) (@imunbiased). Be excellent to each other. Up a WV holler..or in NoVA

3) Today several crucial MoUs were signed that will boost India-France friendship.

- Negative tweets

1) Any negative polls are fake news, just like the CNN, #DonaldTrump

2) Can Hillary please hire the genius/magician who dressed Palin in 2008 and stop dressing like my weird cat-lady aunt who works at JCPenney?— kara vallow (@teenagesleuth)

3) Sasha and Malia Obama, daughters have some selfie fun during the Inaugural Parade for their father President Obama on ... Follow @JessicaDurando

From the examples above, it is clear that tweets can contain valuable information expressing opinions on any topic. However, they may also include specific characters that are not helpful in detecting sentiment polarity; hence, it makes sense to preprocess tweets. This second step consists of converting all tweet texts to lower case. In addition, tweets should be cleaned by removing URLs, hashtag characters (such as #Trump) or user mentions (such as @Trump) as Twitter sentiment-analysis methods are not concerned with these characters. The preprocessing step includes filtering out stop words that are considered unusual discriminant features [11].

After preprocessing, predictions are performed. In this phase, various prediction algorithms, such as the SVM, Bayesian classifier, and Entropy Classifier, can be used to decide the sentiment polarity of tweets. For example, Vishal *et al.* [17] reviewed current procedures for opinion mining such as machine learning and vocabulary-based methodologies. Utilizing different machine learning algorithms like NB, Max Entropy, and SVM, Vishal *et al.* [17] additionally described general difficulties and utilizations of Twitter sentiment analysis.

Go and L.Huang [26] proposed an answer for conclusion examination for Twitter information by utilizing far off supervision, in which their preparation information comprised of tweets with emojis which filled in as uproarious names. Go et al [26] introduced a method to classify the sentiment of tweets. The idea behind it was to aggregate feedback automatically. The sentiment problem was treated as a binary classification, in which tweets were classified into positive and negative. Training data containing tweets with emoticons were collected based on supervision approach that was proposed by Read [27]. To achieve this, Go et al [26] utilized the Twitter API to extract tweets that included emoticons. These were used to identify tweets as either negative or positive. Retweeted posts and repeated tweets were removed. In addition, tweets containing positive and negative emotions were filtered out. Various classifiers such as the NB, MaxEnt, and SVM were employed to classify tweets. Different features were extracted such as unigrams, bigrams, unigrams with bigrams, and unigrams with POS. The best results were obtained by the MaxEnt classifier in conjunction with unigram and bigram features, which achieved an accuracy of 83% compared to the NB with a classification accuracy of 82.7%.

Malhar and Ram [28] proposed the supervised method to categorize Twitter data. The results of this experiment demonstrated that the SVM performed better than other classifiers and, using a hybrid feature selection, achieved an accuracy of 88%. The experiment attempted to combine principal component analysis (PCA) alongside the SVM classifier to reduce feature dimensionality. Furthermore, unigram, bigram, hybrid (unigram and bigram) feature-extraction methods were used. Malhar and Ram [28] showed that integrating PCA with the SVM with a hybrid feature selection could help in reducing feature dimensions and the results obtained a classification accuracy of 92%.

Anton and Andrey [29] developed a model to extract sentiment polarity from Twitter data. The features extracted were words containing n-grams and emoticons. The experiment carried out demonstrated that the SVM performed better than the Naïve Bayes. The best overall performing method was the SVM in combination with unigram feature extraction, achieving a precision accuracy of 81% and a recall accuracy of 74%.

Po-Wei Liang *et al.* [30] designed a framework called an "opinion miner" that automatically investigated and detected the sentiments of social media messages. Annotated tweets were combined for the undertaking of the analysis and in this framework, messages which contained feelings were extracted (non-opinion tweets were removed) and their polarities determined (i.e. positive or negative). To achieve this, the experimenters [30] classified the tweets into "opinion" and "non-opinion" using the NB classifier with a unigram. They likewise disposed of irrelevant features by utilizing the Mutual Information and chi-square extraction strategy. The experimental outcomes confirmed the adequacy of the framework for sentiment analysis in genuine microblogging applications.

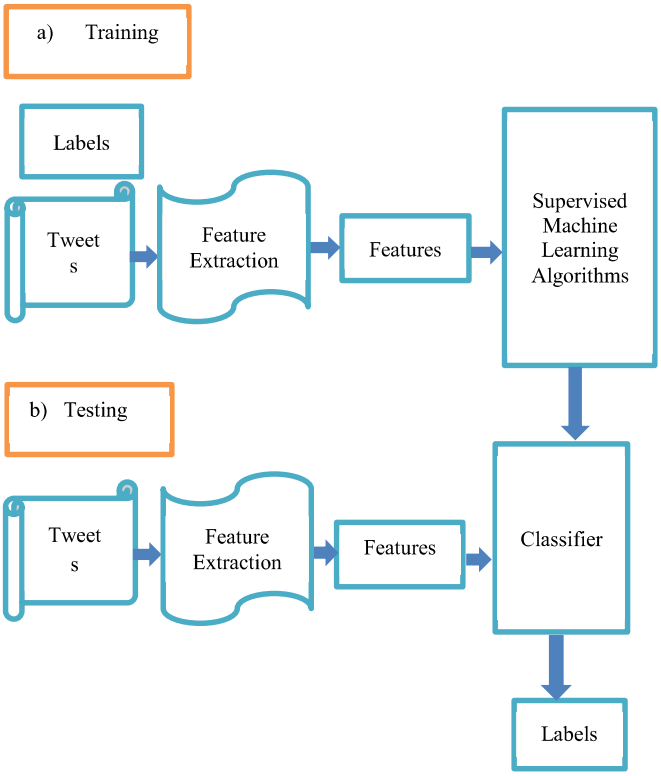

Fig 2. Sentiment Analysis using Supervised Machine Learning Algorithms.

Pak and Paroubek [31] used Twitter API and emoticons to collect negative and positive sentiments, in the same way as Go *et al.* [26]. Sentiment analysis was treated as multi-label, with tweets classified as positive, negative, or neutral. The statistical-linguistic analysis was performed on the collected training data based on determining the frequency distribution of words. The collected training datasets were used to build a classifier and experiments were conducted on the SVM, conditional random fields (CRF), and Multinomial Naïve Bayes (MNB) classifier with different feature selection methods. The MNB with part of speech tags and n-gram features was the technique that produced the best performance in the experiments.

Kouloumpis *et al.* [32] explored the usefulness of various linguistic features for mining the sentiments of Twitter data. The hash-tagged (HASH) and emoticon (EMOT) datasets were utilized to train the classifiers and the iSieve dataset was used for the evaluation. In this study, various feature sets were introduced using unigrams, bigrams, lexicons, micro-blogging and part-of-speech elements. The AdaBoost classifier was trained using these selected features in different combinations. The results showed that part-of-speech features were poor for sentiment analysis of Twitter data whilst micro-blogging features were the most useful. The best results were achieved when n-gram features were employed alongside lexicon and micro-blogging features. An F-score of 0.68 was obtained with HASH datasets and an F-score of 0.65 with HASH and EMOT datasets combined.

Saif *et al.* [33] introduced the idea of merging semantic with unigram and part of speech features. Semantic features are concepts that encapsulate entities mined from Twitter data. The extracted features were used to compute the correlation of entity groups augmented by their sentiment polarities. It is worth noting that incorporating semantic features into the analysis can help in detecting the sentiment of tweets that include entities. Saif *et al.* [33] used three datasets collected from Twitter to evaluate the impact of adding semantic features. In the conducted experiment, the Naïve Bayes classifier was used alongside the extracted semantic features. The findings demonstrated that semantic features led to improvements in detecting sentiments compared to the unigram and part-of-speech features. Nevertheless, for the HCR and OMD datasets, the sentiment-topic approach tended to perform better than the semantic approach. For the HCR, the former achieved an F1 score of 68.15 compared to an F1 score of 66.10 obtained by the semantic approach. For the OMD dataset, an F1 score of 78.20 was reached using the sentiment-topic approach compared to an F1 score of 77.85 achieved by the semantic approach.

Hamdan *et al.* [34] extracted different types of features with the intention of enhancing the accuracy of sentiment classification. Unigram features were introduced as a baseline whereas words were considered independent features. Domain-specific features were also included, such as the number of re-tweets. DBpedia was utilized to mine the concepts contained in tweets; these will be termed DBpedia features. WordNet was used to identify the synonyms of nouns, verbs, adverbs, and adjectives. SentiWordNet was employed to compute the frequency of positive and negative words appearing in tweets and the polarities of these tweets. The experiments showed that adding adjectives, SentiWordNet and DBpedia features led to minor improvements in the accuracy of both the SVM and NB. The ratios of these slight improvements were approximately 2% with the SVM and 4% with the NB.

Akba *et al.* [35] employed feature selection based on information gain and chi-square metrics to elect the most informative features after the stemming and lemmatization processes. The conducted experiments showed that incorporating feature selection metrics with the SVM classifier led to improvements over previous studies. In addition, Saif *et al.* [36] investigated the impact of information gain as a feature selection criterion in order to rank unigrams and semantic features. They concluded that the performance of a classifier can be acceptable even when selecting few distinctive sentiment-topic features using information gain.

B. Twitter Sentiment Analysis using Ensemble Approaches

The basic principle of ensemble methods is to combine multiple classifiers with a view to obtaining more precise and accurate predictions. Ensemble methods are widely used for text classification purposes and in the field of Twitter sentiment analysis, such methods may be advantageous for improving the classification accuracy of Twitter posts.

Xia *et al.* [1] investigated the effectiveness of creating ensemble learners for sentiment classification purposes. The intention was to efficiently mix diverse feature sets and various classification algorithms to create a more powerful classifier. They utilized a gathering system for sentiment classification which was acquired by combining different capabilities and arrangement procedures. Traditional text classification approaches are not suited to sentiment classification as the bag of words (BOW) misses-out some word information. In this study, two feature types (POS and Word-relations) and three classifiers (NB, MaxEnt and SVM) were utilized. Three kinds of ensemble classifiers were proposed and evaluated: namely weighted grouping, fixed grouping, and meta-classifier grouping. The results showed that the ensemble methods led to clear improvements compared to the individual classifier. Moreover, the outcomes proved that the ensemble of both various classifiers with different feature sets produced very significant improvements.

Lin and Kolcz [37] proposed incorporating multiple classifiers into large-scale twitter data. They attempted to train logistic regression (LR) classifiers from the hashed 4-grams as features. The training dataset varied from one to 100 million of examples with ensembles of 3 to 41 classifiers. The experiment showed that the accuracy of sentiment analysis of Twitter data using multiple classifiers was greater than with a single classifier. The drawback of the ensemble method was that the running time increased as n classifiers require n separate predictions. The best performance was obtained when the number of classifiers was 21 and the number of instances was 100 million, achieving a classification accuracy of 0.81.

da Silva *et al.* [38] suggested an ensemble model that consisted of four base classifiers: the SVM, MNB, random forest, and logistic regression. Two approaches were used to represent the features: BOW and feature hashing. The results

gathered illustrated that the ensemble classifier with a combination of BOW and lexicon features led to improvements in the classification accuracy [38]. The ensemble method proposed in [38] attained accuracy scores of 76.99, 81.06, 84.89, and 76.81 for HCR, STS, Sanders, and OMD datasets, respectively.

Hagen, Matthias *et al.* [39] reproduced and combined four Twitter sentiment classifiers to create an ensemble model called "Webis". The impetus behind producing this combination was to utilize the strength of the four classifiers as each one corresponds to different feature sets. Instead of taking the majority vote on predictions from the participated classifiers, Hagen, Matthias *et al.* [39] introduced a confidence score for the four classifiers in order to obtain the final predictions. In their work, they computed the confidence scores for each individual classifier and each class. The classification decisions were made based on the highest scores on average. The Webis classifier was used as a strong baseline because it was the winner in the SemEval-2015 Task 10. The ensemble method produced an F-score of 64.84 for subtask B.

Martınez-Cámara, Eugenio *et al.* [40] employed an ensemble classifier that used various Twitter sentiment approaches to enhance the performance and efficiency of classifying the polarity of tweets. Their model was a combination of a ranking algorithm and skip-gram scorer, Word2Vec, and a linguistic resources-based approach [40]. It is important to highlight that their proposed ensemble method relied upon voting strategies. For evaluating the proposed approach, the training data of the TASS competition were chosen. The results of the experiments showed that a slight improvement was obtained with the ensemble method compared to the ranking algorithm and skip gram methods. The Macro-F1 score achieved by the former was 62.98% compared to a macro F1 score of 61.60% obtained by the latter combination.

Chalothorn and Ellman [41] demonstrated that the ensemble model could produce superior accuracy of emotion classification compared to a single classifier. They [41] combined BOW and lexicon features in the context of ensemble classification and conducted experiments showing that when the extracted features were used in combination with these features, the accuracy of classification increased. The mixture of the SVM, SentiStrength and stacking methods using majority voting produced an F-score of 86.05%; this was considered the highest score.

Fouad *et al.* [42] proposed a system of classifying tweets based on the majority voting of three classifiers: the SVM, NB, and LR. The collected tweets were split into two sets: training and testing. Individual classifiers received the same training set to record their decisions. The ensemble method produced a final decision based on the majority votes collected from the classifiers. The most interesting aspect of their study [42] was that information gain was utilized to reduce the dimensionality of feature vectors. In their work [42], experiments were carried out to examine the impact of information gain on the accuracy of the classifier and the results demonstrated improvements in classification accuracy after feature vector dimensionality was reduced using information gain. Information gain showed clear improvements in the classification accuracies of all datasets. The ratio of improvement was around 15% on average. The results further showed that the proposed majority-voting ensemble classifier achieved an accuracy score of 93.94 compared to a score of 92.71 achieved by the SVM for Sanders datasets. In addition, the majority-voting ensemble classifier achieved an accuracy score of 78.70 compared to 78.10 obtained by the SVM for the Stanford-1K dataset. However, for the HCR dataset, the NB achieved an accuracy score of 85.09 compared to the ensemble methods that obtained a score of 84.75.

C. Twitter Sentiment Analysis using Lexicon based Approaches (Unsupervised Methods)

Normally, lexicon-based methodologies for sentiment analysis depend on the understanding that the polarity of a text sample can be acquired on the grounds of the polarity of the words which comprise it. However, because of the complexity of natural languages, such a basic approach will likely be inadequate since numerous aspects of the language (e.g. the nearness of the negation) are not taken into consideration. As a result, Musto [43] proposed a lexicon-based approach to identify the sentiment of any given tweet T, which began by breaking down the tweet into a number of small-scale phrases, such as $m_1 \dots m_n$ as indicated by the part signs occurring in the content. Punctuations, adverbs and conjunctions constituted the part signal and, at whatever point a part signal occurred in the text, another micro-phrase is constructed.

The sentiment of a tweet was determined by adding the polarity of each smaller micro-phrase after the splitting phase. At this point, the score was standardized across the length of the entire Tweet. In this situation, the micro-phrases were simply exploited to reverse the polarity when a negation was found in the content.

The polarity of a micro-blog post depended on the polarity of the micro phrases which united it:

$$\text{pol}(\text{Tweet}) = \sum_{i=1}^{k} \text{pol}(m_i) \quad \text{and Tweet} = m_1, m_2 - - - - - - m_k \tag{3}$$

The polarity of a micro-phrase (m) depended on the polarity of the terms which composed it:

$$\text{pol}(m_i) = \sum_{j=1}^{n} \text{score}(t_j) \tag{4}$$

The score of each micro-phrase was normalized according to its length

$$\text{pol}(m_i) = \sum_{j=1}^{n} \text{score}(t_j) / m_i \tag{5}$$

Specific POS categories were provided with higher-weight categories including adverbs, verbs, adjectives and valence shifters (intensifiers and down-toners). Several weights were evaluated as follows:

- Emphasized version

$$\text{pol}(m_i) = \sum_{j=1}^{n} \text{score}(t_j) * w_j \tag{6}$$

- Normalized-Emphasized version

$$\text{pol}(m_i) = \sum_{j=1}^{n} \left(\frac{\text{score}(t_j)}{m_i}\right) * w_j \tag{7}$$

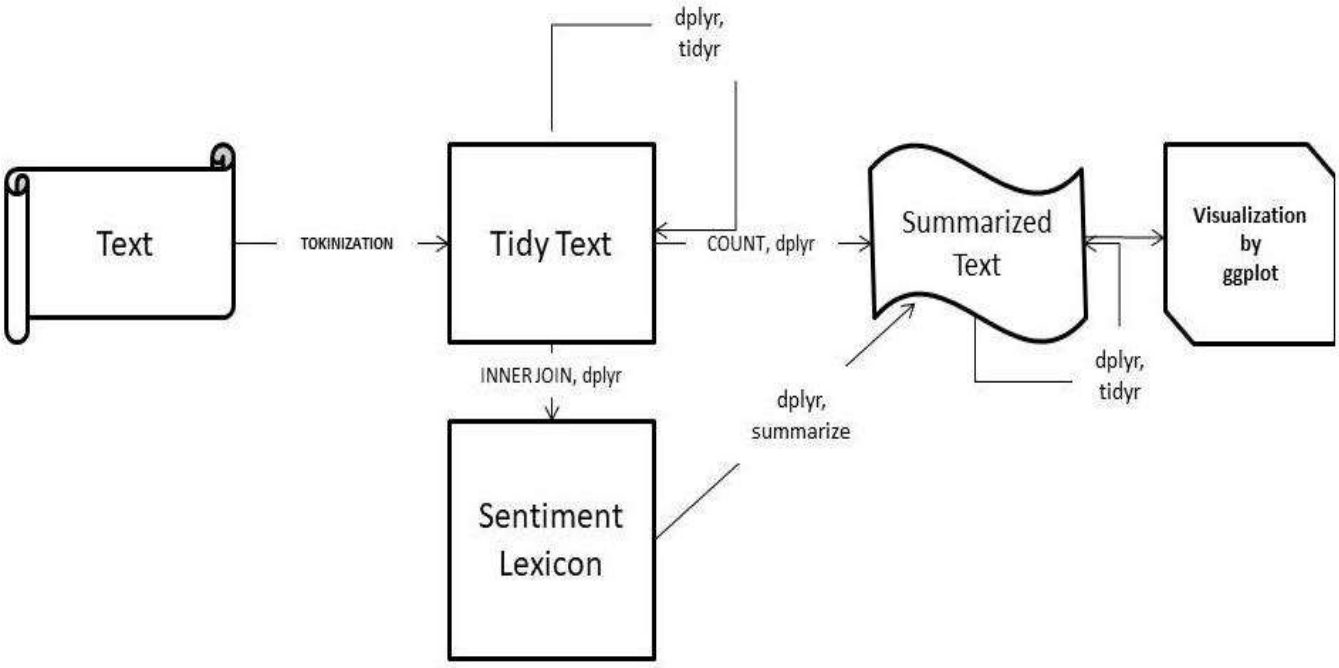

Fig 3. The Estimation Computation Procedure [44].

Lexicon and external lexical resources are SentiWordNet, MPQA and WordNet-Affect, SenticNet are required to compute the $\text{score}(t_j)$. The procedure for the estimation computation is schematically shown in Fig. 3 and can be depicted with the accompanying advances: Lexicon based strategies like the ones we are examining locate the total sentiment of a bit(piece) of content by including the individual sentiment scores for each word in the text [43]. SentiWordNet and MPQA [11] are the most utilized dictionaries that are widely utilized for detecting the sentiment of the given tweets.

According to Xia *et al.* [45], it was an easy task to gather a vast number of unlabeled data from social networks; however, detecting the sentiment labels of these data was very costly. Thus, it was necessary to use unsupervised sentiment analysis approaches. Moreover, unsupervised learning methods are increasingly considered vital as the volume of unlabeled information in social media increases.

Xia *et al.* [45] exploited emotional signals to detect sentiments appearing in social media data. These emotional signals were defined as any information that correlated or was associated with sentiment polarities. Xia *et al.* [45] proposed a framework: Emotional Signals for unsupervised Sentiment Analysis (ESSA). They then proposed modelling emotional indicator to detect the sentiment polarity of posts and to bring this closer to the emotional indictors within the post. Moreover, they proposed modelling word-level emotional indicators to detect the polarity of posts and to bring the polarity of the words closer to the word-level emotional indicators. Stanford Twitter sentiment (STS) and OMD were used as datasets for the conducted experiments. The ESSA framework obtained classification accuracies of 0.726 for the STS and 0.692 for the OMD datasets. The results demonstrated the usefulness of the ESSA framework compared to other techniques.

Azzouza, Noureddine *et al.* [46] presented a real-time architecture to detect opinions in Twitter data. Their system relied on an unsupervised machine learning technique to explore tweets and detect their polarity. This classification technique used a dictionary-based approach to identify the polarity of tweeted opinions and their architecture [46] consisted of multiple modules. Tweets were collected using a tweet-acquisition module that was connected with the Twitter API to retrieve tweets using queries posed. Text was tokenized using a separate module. Then, lexical correction, token standardization, and syntactic correctness were various stages in the tweet-processing module. The researchers introduced an opinion-analysis module to compute the opinion value for emoticons, words, and the average of opinion values. The experiments were conducted based on the SemEval dataset to measure the quality of the real-time architecture. For the SemEval-2013 dataset, the proposed system reached an accuracy score of 0.559 compared to 0.50 obtained by the SSA-UO system proposed by Ortega *et al.* [47]. Furthermore, the architecture proposed in [46] achieved an accuracy of 0.533 compared to 0.539 obtained by the GTI research group for the SemEval-2016 dataset.

Paltoglou and Thelwall [48] employed a lexicon-based approach to estimate the level of emotional intensity to make predictions. This approach was appropriate for detection of subjective texts expressing opinion and for sentiment polarity classification to decide whether the given text was positive or negative. The proposed lexicon-based method achieved F1 scores of 76.2, 80.6, and 86.5 for the Digg, MySpace, and Twitter datasets outperforming all supervised classifiers.

Masud *et al.* [49] applied a vocabulary-based system for sentiment classification, which characterized tweets as positive, negative, or unbiased. This system [49] distinguished and

scored slang utilized in tweets. The experimental outcomes demonstrated that the proposed framework outperformed existing frameworks, accomplishing 92% precision in double characterization and 87% in multi-class grouping. The framework needed to enhance accuracy in negative cases and to review in nonpartisan cases.

Asghar *et al.* [50] proposed an improved lexicon-based sentiment classification that incorporated a rule-based classifier. It aimed to reduce data sparseness and to improve the accuracy of sentiment classification. Classifiers, such as those using emoticons or modifier-negation, or those which were SWN-based or domain-specific, were incorporated sequentially to classify tweets accurately based on their sentiment polarities. The proposed the technique achieved F1 scores of 0.8, 0.795, and 0.855 for three drug, car, and hotel reviews datasets respectively.

D. Twitter Sentiment Analysis using Hybrid Methods

Balage Filho and Pardo [51] introduced a hybrid system for detecting the sentiments present in tweets. Moreover, their system combined three classification methods: machine learning, rule-based, and lexicon-based. Balage Filho and Pardo [51] used the SentiStrength lexicon and the SVM classifier as a machine learning method. The results obtained from the experiments showed that a hybrid system outperformed the individual classifiers, achieving an F-measure of 0.56 compared to 0.14, 0.448, and 0.49 obtained by the rule-based, lexicon-based, and SVM classifiers respectively.

Another hybrid method was proposed by Ghiassi *et al.* [52] who utilized Twitter API to collect tweets. They attempted to combine n-gram features with a developed dynamic artificial neural network (DAN2) sentiment analysis method. Unigram, bigram, and trigram features were identified. Ghiassi *et al.* [52] developed a reduced Twitter lexicon that was used alongside sentiment classification methods. DAN2 and SVM classification models were trained to detect the sentiment of tweets. The collected results showed that the DAN2 learning method performed slightly better than the SVM classifier even when incorporating the same Twitter-specific lexicon. For the negative class, the DAN2 achieved an accuracy of 92.5 on average compared to the SVM, which achieved an accuracy of 91.45. For the positive class, the DAN2 obtained a classification accuracy of 68.2 on average compared to the SVM, which achieved an accuracy of 67.6.

Khan *et al.* [53] proposed a Twitter opinion mining (TOM) framework for tweets sentiment classification. The proposed hybrid scheme in [53] consisted of SentiWordNet analysis, emoticon analysis, and an enhanced polarity classifier. The proposed classifier mitigated the sparsity problems by employing various pre-processing and multiple sentiment methods. The experiments were conducted using six datasets demonstrated that the proposed algorithm achieved an average harmonic mean of 83.3%.

Recently, Zainuddin *et al.* [54] proposed an aspect-based sentiment analysis (ABSA) framework, which contained two principal tasks. The first task used aspect-based feature extraction to identify aspects of entities and the second task used aspect-based sentiment classification. HCTS, STS, and STC datasets were used to evaluate the performance of the proposed hybrid model. This model incorporated rules after mining them with feature extraction methods. Single and multi-word aspects were identified based on a rule-mining technique with heuristic combination in POS patterns. Moreover, the Stanford dependency parser (SDP) was used to detect dependencies between aspects and opinions. Principal component analysis (PCA), latent semantic analysis (LSA), and random projection (RP) feature selection methods were also adopted in the experiments. The new hybrid model combining the ABSA framework, SentiWordNet lexicons, PCA, and the SVM classifier outperformed the existing baseline for sentiment classifications. A classification accuracy of 76.55 was achieved for the STS dataset; 71.62 for the HCTS dataset; as well as an accuracy of 74.24 for the STC dataset.

Asghar *et al.* [55] proposed a hybrid Twitter sentiment system that incorporated four classifiers: a slang classifier (SC), an emoticon classifier (EC), a general purpose sentiment classifier (GPSC), and an improved domain specific classifier (IDSC). Their technique was inspired by the previous studies by Khan *et al.* [53] and Asghar *et al.* [50], which classified tweets using multiple supervised and unsupervised classification models. The proposed framework identified the sentiment of tweets after detecting the presence of slang and emoticons. The results showed that computing the sentiment score of slang expressions lead to an improved accuracy in the sentiment classification of tweets. In terms of studying the impact of SC, the framework proposed by Asghar *et al.* [55] achieved an F-score of 0.92 compared to 0.85 obtained by Masud *et al.* [49]. The results also showed that the presence of emoticons in Twitter sentiment increased classification accuracy from 79% to 85%.

VII. Discussion and Findings

In this section of the study, an attempted was made to compare the different techniques and outcomes of algorithms performance. Table 1 summarizes various supervised machine learning approaches for Twitter sentiment analysis. It is important to mention that the unigram-based SVM is normally considered a benchmark against which the proposed strategies are measured and compared [11]. From Table 1, it is clear that integrating multiple features led to improvements in classification accuracy, especially combining unigrams and bigrams as demonstrated in Go *et al.* [26] and Malhar and Ram [28]. In contrast, Anton and Andrey [29] demonstrated that the SVM classifier when combined with unigram features outperformed hybrid features. According to Saif *et al.* [33], the results showed that incorporating semantic with unigram features produced better performance than the baseline feature selection.

In a similar way, Hamdan *et al.* [34] showed that adding more features such as DBpedia, WordNet and SentiWordNet led to improvements in sentiment classification accuracy. According to Vishal *et al.* [17], machine learning methodologies like NB, Max Entropy, and SVM performed slightly better with bigram features compared to other feature models such as unigrams or trigrams.

TABLE I. THE SUPERVISED MACHINE LEARNING APPROACH FOR TWITTER SENTIMENT ANALYSIS

Study	Methods	Algorithms	Features	Datasets	Outcomes
Go et al [26]	Supervised ML	NB, MaxEnt, and SVM classifiers.	Unigrams, bigrams, POS	Tweets collected using Twitter API	The MaxEnt with both unigrams and bigrams achieved an accuracy of 83% compared to the NB with an accuracy of 82.7%.
Malhar and Ram [28]	Supervised ML	NB, SVM, MaxEnt, and ANN classifiers.	Unigrams, bigrams, hybrids (unigrams+ bigrams)	Tweets collected using Twitter API	The SVM using the hybrid feature selection achieved an accuracy of 88%. In addition, the SVM with the PCA achieved an accuracy of 92%.
Anton and Andrey [29]	Supervised ML	NB and SVM classifiers	Unigrams, bigrams, hybrids (unigrams+ bigrams)	Tweets collected with the online system Sentiment140	The SVM with unigrams reached a precision score of 81% and a recall score of 74%.
Pak and Paroubek [31]	Supervised ML	Multinomial NB and SVM classifiers	Unigrams, bigrams, trigrams	Tweets collected using Twitter API	Multinomial NB with bigrams achieved a better performance compared to unigrams and trigrams.
Kouloumpis et al. [32]	Supervised ML	AdaBoost classifier.	Unigrams, bigrams, lexicon, POS features, and micro-blogging features	The hash-tagged (HASH) and emoticon (EMOT) as training datasets.	An F-measure of 0.68 was achieved for HASH. In addition, an F-measure of 0.65 was obtained by AdaBoost for HASH and EMOT datasets with a combination of n-grams, lexicons and microblogging features
Saif et al. [33]	Supervised ML	NB	Unigrams, POS features, sentiment-topic features semantic features	STS, HCR and OMD datasets	Semantic features outperformed unigrams and POS. However, the sentiment-topic approach performed marginally better than the semantic approach in the case of the HCR and OMD datasets.
Hamdan et al. [34]	Supervised ML	NB, SVM	Unigrams, DBpedia wordNet, and SentiWordNet	SemEval- 2013 datasets	Experiments showed that adding features such as DBpedia, WordNet and SentiWordNet led to a slight increase in the F-measure accuracy. The ratio of these slight improvements was about 2% with the SVM and 4% with the NB.

Table 2 illustrates various ensemble approaches for Twitter sentiment analysis. For the HCR dataset, the ensemble methods proposed by da Silva *et al.* [38] that incorporated LR, RF, SVM, and MNB alongside BOW and lexicon features achieved the F1 score of 76.99. In comparison, Fouad *et al.* [42] showed that the majority voting ensemble method with information-gain feature selection method achieved an accuracy of 84.75. This demonstrates that the ensemble methods proposed by Fouad *et al.* [42] outperformed the ensemble method proposed by da Silva *et al.* [38]. This was due to incorporating the information gain as a feature selection method.

Saif *et al.* [33] showed that the NB classifier achieved an F1 score of 68.15 for the HCR dataset. In comparison to the ensemble methods proposed by da Silva *et al.* [38] which incorporated LR, RF, the SVM, and the MNB attained an F1 score of 63.75 for the HCR dataset. Furthermore, da Silva *et al.* [38] obtained a slight improvement using the MNB with the BOW and lexicon features, producing an F1 score of 68.20 compared to 68.15 obtained by the NB classifier proposed by Saif *et al.* [33].

According to Fouad *et al.* [42], the performance of their ensemble method was marginally better than the SVM classifier for the Sanders dataset, as shown in Table 2. This was attributed to the majority voting idea that was employed to determine the final sentiments of tweets. However, for the HCR dataset, NB with an information gain feature selection achieved the highest accuracy score of 85.09 compared to both the ensemble method proposed by Fouad *et al.* [42] and to the method proposed by da Silva *et al.* [38] producing a score of 76.99.

TABLE II. ENSEMBLE APPROACHES FOR TWITTER SENTIMENT ANALYSIS

Study	Methods	Algorithms	Features	Datasets	Outcomes
Lin and Kolcz [37]	Ensemble	Logistic regression classifier	Hashed byte 4-grams	Large-scale datasets	For 100 million instances, the ensemble methods achieved an accuracy score of 0.81 when the number of classifiers was 21.
da Silva et al.[38]	Ensemble	MNB, RF, SVM, and LR	BOW, lexicon, and feature hashing	Stanford (STS), Sanders, OMD, and HCR datasets	An ensemble classifier achieved higher accuracies when both BOW and lexicon features were utilized. The proposed method achieved accuracy scores of 76.99, 81.06, 84.89 , and 76.81 for HCR, STS, Sanders, and OMD datasets, respectively.
Hagen, Matthias, et al. [39]	Ensemble	NRC, GU-MLT-LT, KLUE, and TeamX	n-grams, ALLCAPS, parts of speech, polarity dictionaries, punctuation marks, emoticons, word lengthening, clustering, negation, stems	SemEval-2013 training	The ensemble method attained an F-score of 64.84 for subtask B in the SemEval-2015 Competition (Task 10).
Martınez-Cámara, Eugenio, et al.[40]	Ensemble	The ranking algorithm and skip-gram scorer, Word2Vec, and linguistic resources-based approach	The ranking algorithm and skip-gram scorer	General Corpus of the TASS competition	The ensemble method achieved a macro F1-score of 62.98%. However, the ranking algorithm and skip-gram obtained a macro F1 score of 61.60%.
Chalothorn and Ellman [41]	Ensemble	The majority vote, SVM, NB, SentiStrength and Stacking.	Sentiment lexicons and BOW features	SemEval-2013	The ensemble classifier obtained an F-score of 86.05% for task 2A.
Fouad et al. [42]	Ensemble	SVM, NB, and LR	Various combinations of BOW, lexicon-based features, emoticon-based and POS features.	Stanford (STS), Sanders, and HCR	For the Sanders datasets, the ensemble (majority voting) classifier achieved an accuracy score of 93.94 compared to 92.71 achieved by the SVM. For Stanford -1K dataset, the majority voting ensemble classifier achieved an accuracy score of 78.70 to 78.10 obtained by the SVM. For the HCR, the NB achieved an accuracy score of 85.09 compared to the proposed majority vote ensemble methods which obtained a score of 84.75.

Table 3 summarizes various lexicon-based algorithm investigated in this paper. Xia *et. al* [45] showed that their lexicon-based sentiment method achieved a classification accuracy of 0.692 for the OMD dataset compared to a classification accuracy score of 76.81 that attained by the ensemble method proposed by da Silva *et al.* [38]. This may attribute to the utilization of the majority voting ensemble classifier and combining lexicons with BOW features.

Table 4 shows the hybrid algorithms explored in this survey. The method proposed by Zainuddin *et al.* [54] obtained an accuracy score of 76.55 % for the STS dataset and outperformed the lexicon-based methods proposed by Xia *et. al* [45] which achieved an accuracy score of 72.6% for the same data set. In addition, the majority-voting ensemble method proposed by Fouad *et al.* [42] achieved a score of 78.70%. The best results were attained by da Silva *et al.* [38] as their ensemble methods scored an accuracy of 81.06% for the STS dataset.

TABLE III. LEXICON-BASED METHODS FOR TWITTER SENTIMENT ANALYSIS

Study	Methods	Algorithms	Features	Datasets	Outcomes
Xia et. al [45]	Unsupervised method (lexicon-based)	Exploring slang sentiment words in Sentiment analysis (ESSA)	Unigrams	STS and OMD datasets	Classification accuracies of 0.726 for the STS dataset and 0.692 for the OMD dataset were achieved.
Azzouza, Nouredd ine, et al. [46]	Unsupervised method		POS features	SemEval-2013, SemEval-2014, SemEval-2015, SemEval-2016	For the SemEval-2013 dataset, the proposed system obtained an accuracy score of 0.559 compared to 0.50 obtained by the SSA-UO. For the SemEval-2016 dataset, the proposed system achieved an accuracy score of 0.533 compared to 0.539 obtained by the GTI.
Paltoglo u and Thelwall [48]	Unsupervised method (lexicon-based)	Emotional lexicon	Unigrams	Digg, MySpace, and Twitter datasets	The proposed lexicon method achieved F1 scores of 76.2, 80.6, and 86.5 for Digg, MySpace, and Twitter datasets, respectively.
Masud et al. [49]	Unsupervised method (lexicon-based)	Lexicon and dictionaries		own datasets	The proposed method of integrating lexicons and dictionaries achieved an accuracy of 92% for binary classification and 87% for multi-class classification.
Asghar et al. [50]	Lexicon-enhanced-Rule-based	Rule-based classifier	Emoticon-handling features and an enhanced feature weighting scheme	Three review datasets	For the second dataset, the proposed technique achieved an F1-measure of 0.795 whilst [56] achieved an F-score of 0.76. For the third dataset, the proposed method achieved an F-score of 0.855 compared to an F-score of 0.77 obtained in [56].

TABLE IV. HYBRID METHODS FOR TWITTER SENTIMENT ANALYSIS

Study	Methods	Algorithms	Features	Datasets	Outcomes
Balage Filho and Pardo [51]	Hybrid	The SVM as the machine learning classifier, and the SentiStrength as the lexicon-based classifier, and the rule-based classifier	BOW	SemEval-2013 Task datasets	The hybrid model achieved an F-score of 0.563 compared to 0.499 obtained by the SVM.
Ghiassi et al.[52]	Hybrid	The Twitter-specific lexicon and DAN2 classifier	Trigrams and bigrams	Own datasets	For the negative class, the DAN2 achieved an accuracy of 92.5 on average compared to 91.45 obtained by the SVM. For the positive class, the DAN2 obtained an accuracy of 68.2 on average compared to an accuracy of 67.6 achieved by the SVM.
Khan et al. [53]	Hybrid	The Enhanced Emoticon Classifier (EEC), Improved Polarity Classifier (IPC), and SentiWordNet Classifier (SWNC)	SentiWordNet Emoticons, sentiment words	Own datasets	An accuracy of 85.7%, precision of 85.3%, and recall of 82.2 recall were achieved.
Zainuddin et al.[54]	Hybrid	Principal component analysis (PCA) and the SVM classifier.	Association rule mining (ARM), POS and Stanford dependency parser (SDP) methods.	STS, HCTS, and STC datasets	The proposed hybrid model outperformed other classifiers for the STS, HCTS, and STC datasets with accuracies of 76.55, 71.62 and 74.24%,respectively.
Asghar et al. [55]	Hybrid	SC, EC, (SentiWordNet), and IDSC classifier.	-	Own datasets	The proposed hybrid classifier achieved an F-score of 0.88 compared to 0.81 achieved by [49].

VIII. Conclusion

In this article, diverse techniques for Twitter sentiment analysis methods were discussed, including machine learning, ensemble approaches and dictionary (lexicon) based approaches. In addition, hybrid and ensemble Twitter sentiment analysis techniques were explored. Research outcomes demonstrated that machine learning techniques; for example, the SVM and MNB produced the greatest precision, especially when multiple features were included. SVM classifiers may be viewed as standard learning strategies, while dictionary (lexicon) based techniques are extremely viable at times, requiring little efforts in the human-marked archive. Machine learning algorithms, such as The Naive Bayes, Maximum Entropy, and SVM, achieved an accuracy of approximately 80% when n-gram and bigram model were utilized. Ensemble and hybrid-based Twitter sentiment analysis algorithms tended to perform better than supervised machine learning techniques, as they were able to achieve a classification accuracy of approximately 85%.

In general, it was expected that ensemble Twitter sentiment-analysis methods would perform better than supervised machine learning algorithms, as they combined multiple classifiers and occasionally various features models. However, hybrid methods also performed well and obtained reasonable classification accuracy scores, since they were able to take advantage of both machine learning classifiers and lexicon-based Twitter sentiment-analysis approaches.

One of the greatest difficulties encountered was in determining the best approach for detecting sentiments in Twitter data because comparing various approaches is a highly challenging task when there is a lack of agreed benchmarks. This difficulty with an absence of well-defined benchmarks was thus addressed in [10] and was found to be mitigated by relying on data sets that had been used for evaluating various algorithms in microblogging sentiment competitions such as SemEval'13 datasets.

Interesting area for future study includes the fluctuations in the performance of sentiment analysis algorithms in cases where multiple features are considered. In other words, combining various features was found to lead to improve the performance in most cases, but substandard performance in others. Thus, an exploration into the causes of these performance instabilities would be an intriguing direction for future works. Another might be to investigate the data sparsity issue using both ensemble and hybrid approaches. The intention behind this is to measure the robustness of various Twitter sentiment approaches the data sparsity. A further area of study might be the utilization of active learning techniques to detect Twitter sentiments and to increase the confidence of decision makers.

References

[1] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," Information Sciences, vol. 181, no. 6, pp. 1138-1152, 2011/03/15/ 2011.

[2] R. Sharma, S. Nigam, and R. Jain, "Opinion mining of movie reviews at document level," arXiv preprint arXiv:1408.3829, 2014.

[3] R. Sharma, S. Nigam, and R. Jain, "Polarity detection at sentence level," International Journal of Computer Applications, vol. 86, no. 11, 2014.

[4] D. Factiva, "Quick Study: Direct Correction Established Between Social Meidia Engagement and Strong Financial Performance," PR News, 2009.

[5] S. R. Das and M. Y. Chen, "Yahoo! for Amazon: Sentiment extraction from small talk on the web," Management science, vol. 53, no. 9, pp. 1375-1388, 2007.

[6] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," Icwsm, vol. 10, no. 1, pp. 178-185, 2010.

[7] I. Twitter, "Second Quarter 2016 Report," ed, 2016.

[8] I. Twitter, "Twitter IPO Prospectus," ed, 2013.

[9] Alexa.com, "Website Traffic Ranking," ed, 2017.

[10] A. DuVander, "Which APIs are handling billions of requests per day?," Programmable Web, 2012.

[11] A. Giachanou and F. Crestani, "Like It or Not: A Survey of Twitter Sentiment Analysis Methods," ACM Comput. Surv., vol. 49, no. 2, pp. 1-41, 2016.

[12] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," Business horizons, vol. 53, no. 1, pp. 59-68, 2010.

[13] A. Abirami and V. Gayathri, "A survey on sentiment analysis methods and approach," in Advanced Computing (ICoAC), 2016 Eighth International Conference on, 2017: IEEE, pp. 72-76.

[14] K. P. Murphy, "Naive bayes classifiers," University of British Columbia, vol. 18, 2006.

[15] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," Comput. Linguist., vol. 22, no. 1, pp. 39-71, 1996.

[16] A. S. Nugroho, A. B. Witarto, and D. Handoko, "Support vector machine," Teori dan Aplikasinya dalam Bioinformatika, Ilmu Komputer. com, Indonesia, 2003.

[17] V. Kharde and P. Sonawane, "Sentiment analysis of twitter data: A survey of techniques," arXiv preprint arXiv:1601.06971, 2016.

[18] A. Harb, M. Plantié, G. Dray, M. Roche, F. Trousset, and P. Poncelet, "Web Opinion Mining: How to extract opinions from blogs?," in Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology, 2008: ACM, pp. 211-217.

[19] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, 2002: Association for Computational Linguistics, pp. 79-86.

[20] J. Khairnar and M. Kinikar, "Machine learning algorithms for opinion mining and sentiment classification," International Journal of Scientific and Research Publications, vol. 3, no. 6, pp. 1-6, 2013.

[21] C. Wu, L. Shen, and X. Wang, "A new method of using contextual information to infer the semantic orientations of context dependent opinions," in Artificial Intelligence and Computational Intelligence, 2009. AICI'09. International Conference on, 2009, vol. 4: IEEE, pp. 274-278.

[22] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," Computational linguistics, vol. 37, no. 2, pp. 267-307, 2011.

[23] T. Zagibalov and J. Carroll, "Unsupervised classification of sentiment and objectivity in Chinese text," in Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I, 2008.

[24] A. Tripathy and S. K. Rath, "Classification of sentiment of reviews using supervised machine learning techniques," International Journal of Rough Sets and Data Analysis (IJRSDA), vol. 4, no. 1, pp. 56-74, 2017.

[25] M. R. Saleh, M. T. Martín-Valdivia, A. Montejo-Ráez, and L. Ureña-López, "Experiments with SVM to classify opinions in different domains," Expert Systems with Applications, vol. 38, no. 12, pp. 14799-14804, 2011.

[26] [26] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Report, Stanford, vol. 1, no. 2009, p. 12, 2009.

[27] J. Read, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification," in Proceedings of the ACL student research workshop, 2005: Association for Computational Linguistics, pp. 43-48.

[28] M. Anjaria and R. M. R. Guddeti, "Influence factor based opinion mining of Twitter data using supervised learning," in 2014 Sixth International Conference on Communication Systems and Networks (COMSNETS), 2014, pp. 1-8.

[29] A. Barhan and A. Shakhomirov, "Methods for Sentiment Analysis of twitter messages," in 12th Conference of FRUCT Association, 2012.

[30] P.-W. Liang and B.-R. Dai, "Opinion mining on social media data," in Mobile Data Management (MDM), 2013 IEEE 14th International Conference on, 2013, vol. 2: IEEE, pp. 91-96.

[31] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in LREc, 2010, vol. 10, no. 2010.

[32] E. Kouloumpis, T. Wilson, and J. D. Moore, "Twitter sentiment analysis: The good the bad and the omg!," Icwsm, vol. 11, no. 538-541, p. 164, 2011.

[33] H. Saif, Y. He, and H. Alani, "Semantic sentiment analysis of twitter," in International semantic web conference, 2012: Springer, pp. 508-524.

[34] H. Hamdan, F. Béchet, and P. Bellot, "Experiments with DBpedia, WordNet and SentiWordNet as resources for sentiment analysis in micro-blogging," in Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), 2013, vol. 2, pp. 455-459.

[35] F. Akba, A. Uçan, E. A. Sezer, and H. Sever, "Assessment of feature selection metrics for sentiment analyses: Turkish movie reviews," in 8th European Conference on Data Mining, 2014, vol. 191, pp. 180-184.

[36] H. Saif, Y. He, and H. Alani, "Alleviating data sparsity for twitter sentiment analysis," 2012: CEUR Workshop Proceedings (CEUR-WS. org).

[37] J. Lin and A. Kolcz, "Large-scale machine learning at twitter," in Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, 2012: ACM, pp. 793-804.

[38] N. F. F. da Silva, E. R. Hruschka, and E. R. Hruschka, "Tweet sentiment analysis with classifier ensembles," Decision Support Systems, vol. 66, pp. 170-179, 2014/10/01/ 2014.

[39] M. Hagen, M. Potthast, M. Büchner, and B. Stein, "Webis: An ensemble for twitter sentiment detection," in Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), 2015, pp. 582-589.

[40] E. Martınez-Cámara, Y. Gutiérrez-Vázquez, J. Fernández, A. Montejo-Ráez, and R. Munoz-Guillena, "Ensemble classifier for Twitter Sentiment Analysis," 2015.

[41] T. Chalothom and J. Ellman, "Simple Approaches of Sentiment Analysis via Ensemble Learning," Berlin, Heidelberg, 2015: Springer Berlin Heidelberg, pp. 631-639.

[42] M. M. Fouad, T. F. Gharib, and A. S. Mashat, "Efficient Twitter Sentiment Analysis System with Feature Selection and lassifier Ensemble," in International Conference on Advanced Machine Learning Technologies and Applications, 2018: Springer, pp. 516-527.

[43] C. Musto, G. Semeraro, and M. Polignano, "A comparison of lexicon-based approaches for sentiment analysis of microblog posts," Information Filtering and Retrieval, vol. 59, 2014.

[44] J. Silge and D. Robinson, Text Mining with R: A Tidy Approach. O'Reilly Media, 2017.

[45] X. Hu, J. Tang, H. Gao, and H. Liu, "Unsupervised sentiment analysis with emotional signals," in Proceedings of the 22nd international conference on World Wide Web, 2013: ACM, pp. 607-618.

[46] N. Azzouza, K. Akli-Astouati, A. Oussalah, and S. A. Bachir, "A real-time Twitter sentiment analysis using an unsupervised method," in Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics, 2017: ACM, p. 15.

[47] R. Ortega, A. Fonseca, and A. Montoyo, "SSA-UO: unsupervised Twitter sentiment analysis," in Second joint conference on lexical and computational semantics (* SEM), 2013, vol. 2, pp. 501-507.

[48] G. Paltoglou and M. Thelwall, "Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 3, no. 4, p. 66, 2012.

[49] F. M. Kundi, A. Khan, S. Ahmad, and M. Z. Asghar, "Lexicon-based sentiment analysis in the social web," Journal of Basic and Applied Scientific Research, vol. 4, no. 6, pp. 238-48, 2014.

[50] M. Z. Asghar, A. Khan, S. Ahmad, M. Qasim, and I. A. Khan, "Lexicon-enhanced sentiment analysis framework using rule-based classification scheme," PloS one, vol. 12, no. 2, p. e0171649, 2017.

[51] P. Balage Filho and T. Pardo, "NILC_USP: A hybrid system for sentiment analysis in twitter messages," in Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), 2013, vol. 2, pp. 568-572.

[52] M. Ghiassi, J. Skinner, and D. Zimbra, "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network," Expert Systems with applications, vol. 40, no. 16, pp. 6266-6282, 2013.

[53] F. H. Khan, S. Bashir, and U. Qamar, "TOM: Twitter opinion mining framework using hybrid classification scheme," Decision Support Systems, vol. 57, pp. 245-257, 2014.

[54] N. Zainuddin, A. Selamat, and R. Ibrahim, "Hybrid sentiment classification on twitter aspect-based sentiment analysis," Applied Intelligence, pp. 1-15, 2017.

[55] M. Z. Asghar, F. M. Kundi, S. Ahmad, A. Khan, and F. Khan, "T-SAF: Twitter sentiment analysis framework using a hybrid classification scheme," Expert Systems, vol. 35, no. 1, 2018.

[56] F. M. Kundi, S. Ahmad, A. Khan, and M. Z. Asghar, "Detection and scoring of internet slangs for sentiment analysis using SentiWordNet," Life Science Journal, vol. 11, no. 9, pp. 66-72, 2014.