A Review of Sentiment Analysis in Social Media Perspectives

Noralhuda N. Alabid
Department of Computer Science
Faculty of Education
Najaf, Iraq
Noralhudan.hadi@uokufa.edu.iq
Orcid.org/0000-0001-5695-1532

DOI: http://dx.doi.org/10.31642/JoKMC/2018/1 I0201
Received Aug. 25, 2023. Accepted for publication Nov. 21, 2023

Abstract— *The widespread use of the Internet and social media platforms has led to an increase in the number of individuals who declare their feelings publicly. Therefore, sentiment analysis systems have proceeded because of their crucial role in determining the personal opinions of users. This is can greatly influence the decision-making process in various fields. To create a robust and reliable sentiment analysis system, it was necessary to apply techniques capable of dealing with these scattered opinions. Natural language processing techniques are commonly used to extract information from unstructured text data published by humans. The comments and posts in social media platforms are often ignore the grammar rules and sentence structure. This is resulting in many ambiguities in lexical, syntactic, and semantic aspects. As a result, researchers have developed different methods for text mining and defining real information. This survey aims to study the different methods used in sentiment analysis filed. We discussed two common models of classification, including the vocabulary-based model and the supervision-based approach.*

Keywords—***Sentiment Analyses; Twitter; Social Media; Pre-processing technique; Supervise Learning algorithme; Unsupervised Learning algorithmes.***

I. INTRODUCTION

The widespread of Web 2.0 technology encourages internet users to contribute more in various online platforms. Internet surfers share their emotions, opinions and thoughts related to different issues via many of social networks such as Twitter, personal Blogs, forums, and etc.

This new technology results in a huge amount of raw data which requires innovative and efficient data mining techniques in order to extract valuable knowledge. Also, these opinions and feelings are related to our lives, thus it is needed to analyze these data to automatically monitor public ideas to help in making decision. For that, the subject of sentiment analysis has obtained more attention over the past decade among the researchers. Since the year of 2004, sentiment analysis research field has grown and become the most active area, with a huge raise in the number of research focused on sentiment analysis.

There are various automatically techniques to extract users opinions on particular issues, events, services, products such as Opinion Mining, text mining and Sentiment Analysis. The main obstacles are these opinions are declared in different forms such as comments, tweets, reviews, etc. and with different linguistic form. Furthermore, most of the comment languages are scattered with incorrect grammar and spell. Thus, analyzing the semantic polarity of these opinions is challenging task.

Sentiment analysis or semantic polarity can be defined as a classifiers used to classify a given texts as positive or negative subjects. Recognizing semantic polarity in a specific text need based on founding some of phrases such as good, professional, excellence, etc. These words are considered key indicators to build a robotic model for sentiment classification. However, these particular polar words mostly are unable on extraction actual sentiment of text without some of other considerations. This is due to that these words can have different polarity in different domains. Thus, when analyzing opinions, it is important to pay attention to analyze contextual and syntactic information [1].

The field of sentiment analysis has gained significant attention from researchers in recent years, resulting in many surveys and review articles. For instance, Ramanathan et al. [2] presented a study discussing general methods for text mining and sentiment analysis, and identified the main issues that need to be addressed in the future. The authors also analysed the problem of opinion detection in the case of fake and spam comments. In another survey, authors in [3] provided a comprehensive overview of sentiment analysis techniques and their applications. They also discussed other areas related to sentiment analysis, such as emotion detection. A survey based

on deep learning approaches for sentiment analysis was presented in [4], while other authors reported on the differences in efficiency between classification techniques for opinion mining in [5]. Birjali et al. [6] conducted a comprehensive investigation of the majority of sentiment analysis techniques.

Several studies utilized different techniques for sentiment analysis extraction such as the one which based on emotional expressions [3], polarity of phrases [4-7], and aspect extraction [8]. In[9], authors apply Part-of-speech (POS) to analyse phrases and define the polarity of texts. Another study considered verbs as a base term that is used for the sentiment classification [10]. Aspect based approach has been used to find the best features, combinations, categorizations, classifications of sentiment analysis [11-12]. There are two main approaches of automatic subjectivity distraction: Machine Learning approaches and Lexicon-based approach [13]. Each of them has powerful and drawback points which are diverse depending on the size of the corpus (texts), subjects and opinion of texts. In this survey, we stated one of a popular subject which is about interpreted and classification textual posts in social media. The paper is organized as follows. The second section discusses the background-of the sentiment-analysis field in many domains. The third section explains the outline needs of sentiment classification. Finally, the conclusion is presented in the fourth section.

II. THE BENEFITS OF SENTIMENT ANALYSIS

The importance of gathering people's opinions has increased with the rapid growth in the number of individuals, services, companies, and government institutions. Most people prefer to use technology in business, communication, finding new information and expressing their feelings [14]. Thus, the application of sentiment analysis is helpful in many areas field. For example, some applications of sentiment analysis are discussed in the following subsections

A. Business domain

Internet users also post comments on many topics, including politics, religion, and social issues. The political organizations have used the opinions of people in social media to identify the general satisfaction of the public about the behavior of the overall policy [15] or monitor people reaction regarding certain policies. For instance, Falck et al.[16]analyzed the effect of political tendencies of newspapers on changing voters decision. These ideas could be used for reporting the political leaders of threats, problems, or potential issues with their community[17].

B. Bolitical domain

Internet users also post comments on many topics, including politics, religion, and social issues. The political organizations have used the opinions of people in social media to identify the general satisfaction of the public about the behavior of the overall policy [15] or monitor people reaction regarding certain policies. For instance, Falck et al.[16]analyzed the effect of political tendencies of newspapers on changing voters decision. These ideas could be used for reporting the political leaders of threats, problems, or potential issues with their community[17].

C. Healthcare domain

The fields of sentiment analysis have taken wide range in health sector. For example, it allowed health care workers to gain information related diseases, drug reactions, and epidemics. Clark et al.[18] analyzed twitter data using each of logistic classifier and neural network model to examine tweets that discussed the experiences of patients who cached the breast cancer. Alabid et al.[19] analyzed tweets data regarding to the covid 19 vaccines for monitoring public reaction. In a study by RamyaSri et al.[20], sentiment analysis was used to analyze patient reviews of hospitals, and the results showed that sentiment analysis can be a useful tool for identifying areas that require improvement in healthcare facilities. In another study by Kohli et al.[21], sentiment analysis was used to analyze Twitter data to monitor public opinion and perception of a new healthcare policy. Furthermore, sentiment analysis can also be used to detect and predict mental health issues. In a study by[22], sentiment analysis was used to analyze the language used in posts published in social media to identify individuals at risk of depression. The study declared that there is some of a specific phrases related of negative sentiment could be associated with a higher risk of depression. This showed that social media can give a good opportunity for patients to explain their needs and concerns. Thus, analyzing these activities on social media based on sentiment analysis application is helpful for deducing a patient's healthcare coverage and determining their treatment needs.

This showed that social media can give a good opportunity for patients to explain their needs and concern. Thus, analyzing these activities on social media based on sentiment analysis application is helpful for inferring a patient's healthcare coverage and determining their treatment needs.

III. SENTIMENT ANALYSIS CHALLENGES

Sentiment analysis has become a crucial tool for understanding public opinion and their thoughts. However, this field has some of challenges:

1. Sentiment analysis system faces the problem of the ambiguity of the language used in social media, in addition to its diversity and the use of diverse meanings across contexts. This makes some of difficulty in interpreting sentiment. [11].

2. Sentiment analysis system must take into account sentiment in each specific field. For example, the feelings used to evaluate products differ from those found in opinions about political positions [21].

3. Sentiment analysis model must consider ethical and privacy issues. Collection and analysis of opinions raise questions about data privacy, agreement, and potential misuse of sentiment data [18].

IV. THE FRAMEWORK OF SENTIMENT ANALYSIS

Sentiment analysis involves a range of problems that requires outfitting variety NLP methods. Sentiment analysis is a research area that focuses on how computers interpret and manipulate human language to create useful applications. Thus, sentiment analysis as one of the areas of information retrieval subjects always contend with NLPs unsolved issues [23], such as Fake new detection [24], topic modeling[25], and text summarization [26].

Generally, NLP is based on logic, mathematics, linguistics, and computer science. It is used in different area such as machine learning, data mining, information retrieval, speech recognition , translations and others [13]. Different NLP methods have been explored for opinion mining and structuring the text. These methods involve: pre-processing step, feature extraction and selection [27]. To outfit the decision, different algorithms can be used such algorithms of machine learning. These techniques are discussed in detail in the following section. Figure 1 explains the general steps of Sentiment Analysis algorithms.

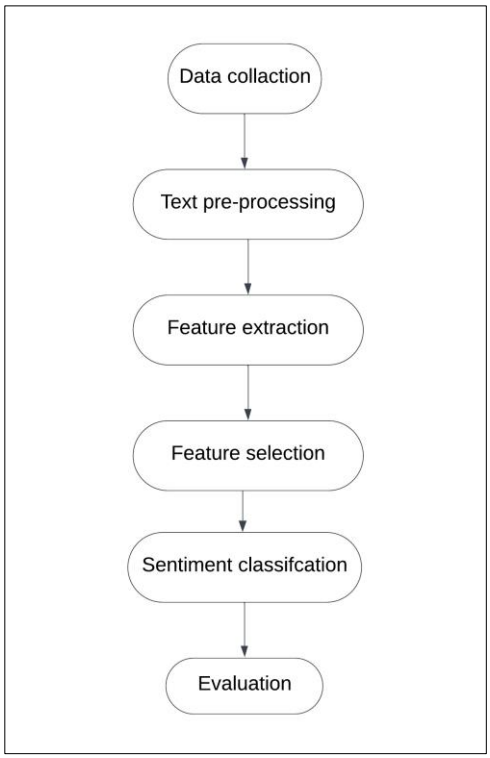

Figure 1.Sentiment Analysis Steps

A. Data acquisition and extraction

1. Data acquisition

Collecting corpuses is performed at the beginning of sentiment analysis processes. There are various sources and tools to obtain it. In general, textual data could be collected from other research, or by searching in the websites. One approach is to collect data from social media platforms, Social media data sources are rich material for investigation and analysis Individual, collective and behavioral life of individuals. Many sites, such as Twitter and Facebook, can be used to collect comments related to the topic of the problem [26]. There is another source which is Forums. It is an online discussion platform where people can ask, share experiences, make social connections, and discuss subjects of common interest [26]. Interview Transcripts is a process by which oral interviews is re-written as text documents. This process can be implemented online or by using a recorded video or audio. Sentiment analysis based on this type of transcripts has been implemented in set of studies [28].

2. Data Extraction

Data extraction is a process that used to transform the collected data to a specific format that can be used for sentiment analysis. This process is varying such as such as removing irrelevant information, improve errors, and standardizing the format of data [29]. In some cases, data extraction could involve employed humans to label data with specific categories. In spite of that this process can improve accuracy but it is time consuming and costly process. Natural language processing (NLP) algorithms such as machine learning and statistical techniques can be used for this process. The accuracy of labelling of these techniques can vary and it depends on the complexity of the text and the quality of the data [30].

B. Pre-processing of textual data

Usually, texts on social media appear in an unstructured form with some useless information. Also, one of the main obstacles in sentiment analysis is the large size of texts (corpus). NLP techniques transfer text from its format to other structure that uses as an input to the sentiment analysis models. Thus, NLP techniques are used to reduce the size of textual data [27]. Figure 2 shows the flowchart of pre-processing techniques commonly used in sentiment analysis.

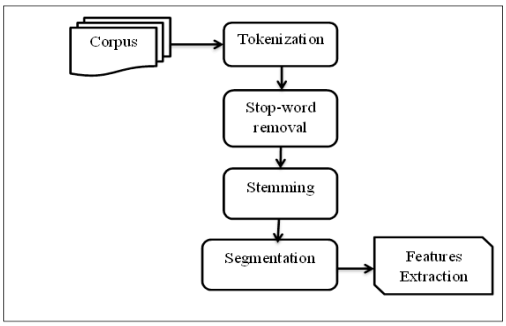

Figure 2.Text Pre-processing Scheme

- In sentiment analysis corpus is the main text that contains the unstructured textual data which could be in different format such as TXT, CSV, XML, or etc.

-Tokenization: it is one of the essential preprocessing techniques. It is responsible for extracting the single words from the main corpus as well as filter out corpus from numbers and punctuations [31]. For example, the sentence "My camera phone has good resolution" would be tokenized to individual words like this "My", "camera", "phone", "has", "good", and "resolution".

-Stop word removal task: this technique focuses on deleting useless words, such as "on", "a", "the", "being", etc. Applying this task decreases the size of the index and the consuming time. However, the stop-word removal task could affect the accuracy of the analysis system. Specifically the misleading could happen when removing the negative phrase such as " not, no, nor" from sentence. However, it is important to be careful when removing some negative phrases that could affect on the real meaning of sentence. [32].

- Stemming is another preprocessing technique used in sentiment analysis. Most English words have several types of morphology in tenses "go- went- gone", nouns and adjectives. A human can easily distinguish between these words but it will be difficult to do that by computers. Thus, the stemming task could be used to solve this obstacle by transferring the tested words to its root. For example, after stemming, the words "intelligent" and "intelligence" are saved as "intellig". Implementing the stemming task reduces redundancy. This will help in decreasing the effort of linking words with lexicon in the more advanced task [3].

-Segmentation: it also plays an essential role in mining web data. This task can be implemented to handle the punctuation marks such as "?" or "!" without ambiguity. However, in case the punctuation marks ".", we need to whether the period marks the end of a sentence or some something else. For example, the period mark could be a part of abbreviations "Dr." or numbers "4.5". However, some solutions could be used to overlapping this ambiguity such as using the Decision Tree method. There are other techniques that could be used to become result of sentiment analysis more accurate and reliable such as include rule-based systems, statistical methods, and machine learning algorithms [33].

C. Features extraction

Feature extraction is a fundamental task in the field of sentiment analysis technology because of its effective role in influencing the results of the classification. This task based on select the most relevant features for the preprocessed dataset to use it as the main input for classifier to define the polarity opinions of text to; positive, negative, or neutral. Many feature extraction methods can be used such as bag of words, n-grams, and word embedding. The bag of words model used the total number of each word in dataset as the feature of attributes. While N-grams method based on count the sequences of n words in the text. Word embedding use neural network techniques to convert each word in text into a numeric vector which can then be used as a feature.

Several algorithms are suggested to evaluate the importance of features by assigning a specific weight for each feature in related text. The most common algorithms are feature frequency (FF) and Term Frequency Inverse Document Frequency (TF-IDF). Feature frequency idea counts the frequency number of appearing each token in documents to find the weight of features. TF-IDF is the most popular method that based on the numerical statistic to evaluate the weight of features in texts. TF-IDF is given by the equations (1)

$$\mathrm{TF}-\mathrm{IDF}=\mathrm{TF}_{(\mathrm{f,d})}*\mathrm{IDF} \qquad (1)$$

Term Frequency (TF) measures how often a feature (f) occurs in a document. It is found by (2).

$$\mathrm{TF}_{(\mathrm{f,d})}=\frac{\mathrm{n}_{\mathrm{f,d}}}{\sum_{\mathrm{w}}\mathrm{n}_{\mathrm{w,d}}} \qquad (2)$$

Where n (f,d) is the number of times the feature f appears in document d, $\sum_{w} n_{w,d}$ is the total number of features in document d.

Inverse document frequency (IDF) is used to measure the importance of features in all documents. It is calculated (3).

$$\mathrm{IDF}_{\mathrm{f}}=\log\frac{|\mathrm{D}|}{\mathrm{d}_{\mathrm{f}}} \qquad (3)$$

Where |D| is the total number of documents in the corpus, df is the total number of documents that contain the detected feature f [32]. Another Features extraction approach is Part-of-Speech (POS). This technique classifies words basically on their grammatical similarity properties like, verbs, pronoun, nouns, preposition, adjectives, and adverbs. For example the following sentence "my friend has a smart Phone" will be labeled based on POS to: My (pronoun), friend (noun), has (verb),a (article) smart (adjective), and phone (noun). Some sentiment analysis research depended on extracting adjectives to identify opinions [34].

D. Feature selection

Feature selection is a process of identifying and eliminating irrelevant or redundant input variables from the feature list, which is used to create a predictive model. The inclusion of such features can lead to increased computational cost and decreased performance of the sentiment analysis process. There are four main categories of feature selection approaches, including filter, wrapper, embedded, and hybrid methods.

- Filter approach is one of the most common techniques when dealing with a high number of features. One of its advantages is that it is less costly compare with other methods. It is based on finding type of link between the feature and the target variable. However, the filter approach does not take into account the interactions between the features. This could lead to detect unsuitable features. To improve performance, filter method is often combined with

other feature methods such as wrapper or embedded approaches, to achieve better performance. Different methods based on filter approach are found such as Mutual information [35], Information Gain[36], and Chi-square [37].

- Wrapper approach: this feature selection method is based on machine learning algorithms. It selects a subset of features and trains a model by using them. Then by using the resulting performance of the applied model, the decision of adding or removing features is derived. Typically. This method is computationally very expensive when dealing with a huge-number of features[38].
    - Embedded methods: Embedded methods perform feature selection during the training step. It takes in consider the interaction between features during the training process. Generally, this approach is mostly used with learning algorithms. In addition, this approach is considered efficient when compared with the wrapper approach. Several types of embedded methods could be used such as word embedded and document embedded. Other examples of embedded methods [39].
    - The hybrid approach is a combination of the filter and wrapper methods. At first this approach uses the filter approach to select set of nominees' features from the original feature set. Then the wrapper approach is used to classify the nominee features. This approach produces the best accuracy by taking the advantage of these two approaches. [40]. The hybrid approach is a computationally efficient method that provides a good balance between the filter and wrapper approaches [41].

V. METHODS OF SENTIMENT ANALYSIS

Sentiment analysis is a rapidly developing research area and has a broad range of applications. Researchers have applied various approaches to improve performance and overcome sentiment analysis challenges. However, selecting the appropriate approach for sentiment analysis is critical and requires careful consideration. In general, sentiment analysis techniques can be classified into two main categories: machine learning approaches and lexicon-based approaches. The following section provides an overview of these methods that are commonly used for conducting sentiment analysis.

A. Machine Learning approaches

Machine learning algorithms are widely used for sentiment analysis to classify texts into negative, positive, or neutral opinions. Two common approaches are suggested to implement sentiment analysis tasks: supervised learning [42] and unsupervised learning approach [43]. It is worth mentioning that the choice of the appropriate machine learning algorithm and feature selection/extraction technique depends on the characteristics of the dataset and the objectives of the sentiment analysis task. Therefore, researchers need to carefully select and tune the algorithms and techniques to achieve the desired performance

1) Supervised learning approach

The main idea of this approach is based on training subset of documents by labeling them into positive, neutral, or negative. The trained data is used to learn the classifier to build the predictive model. Then, the rest of unlabeled data is used to test that model to finally find the best feature set [44]. . There are many classifiers for this purpose such as Naive Bayes, support vector machines (SVM), and artificial neural networks (ANN). In the next paragraph, we discuss the Naive Bayes and SVM classifier. These algorithms have obtained high accuracy rates in sentiment analysis. Their achievement can be further improved by using efficient feature selection and extraction techniques.

In the following paragraphs, we will state two commonly used machine learning algorithms for sentiment analysis: Naive Bayes and Support Vector Machine (SVM) classifier.

1.1) Naive Bayes Classifier

It is the most common classifier that is implemented in the field of text classification. This model is constructed based on Bayes Theorem. This theory assumes that all features of an object are independent and they have no effect on each other. Many studies used Naïve Bayes to classifier texts [45, 46].This model can be obtained by (4).

$$c=\text{argmac}_{c\in C}\frac{P(F_i|C_h)P(C_h)}{P(F_i)} \qquad (4)$$

Where C refers to categories, h assigns to the number of categories, F_i denotes to set of features. Due to the constant value of the P (Fi) through the corpus, it is acceptable to drop it. Therefore, the formula becomes as follows (5):

$$c=\text{argmac}_{c\in C}\ P(F_i|C_h)P(C_h) \qquad (5)$$

To classify documents to their categories, it is needed to compare each feature with all others in features vectors. Mathematically, we can determine the best classifier by (6):

$$O=\text{argmac}_{h\in 1,\ldots\ldots h}\ P(C_h)\prod_{i\in 1}^{n}P(F_i|C_h) \qquad (6)$$

1.2) SVM Classifier

Support Vector Machines (SVM) is a supervised machine learning algorithm used for classification and regression analysis. SVM depends on finding the optimal hyperplane which has to be on the same maximum distance between documents. Furthermore, this technique assumes that finding the longest distance between boundary and documents will reduce the error rate of having an indecisive classification [47]. As shown in Figure 2, there are many different boundaries that can offer many different solutions to the problem. The two boundaries (B1, B2) are classified as a no optimal separator since they are close to documents, while B3 is the perfect one. SVM is known for its high accuracy and

robustness in classification tasks. However, SVM requires careful selection of optimal hyperplane and it can be computationally expensive when working with large datasets. Despite its challenges, SVM has been successfully applied in many area research like sentiment analysis, image classification, and bioinformatics as presented in [48, 39, 41,19].

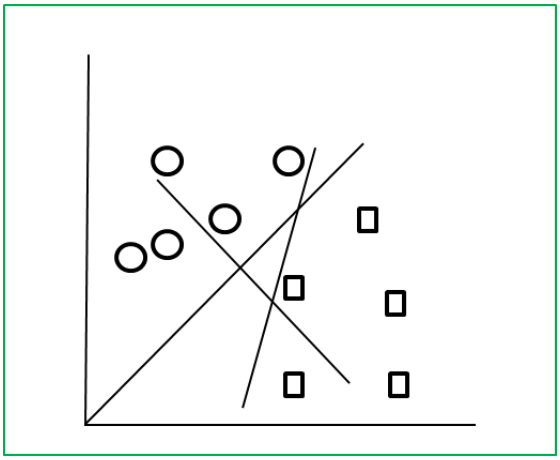

Figure 2: SVM classifyier.

1.3) Artificial Neural Networks

Artificial Neural Networks (ANNs) are computational models based on artificial networks. It has the ability to learn and perform complex tasks. It is widely used in various fields such as natural language processing, computer vision, and speech recognition. Feedforward neural network is consider one of common type of ANN that is used in sentiment analysis. Furthermore, other techniques of neural networks such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have also shown promising results in sentiment analysis [50].

2) Unsupervised learning approach

Unsupervised learning approach is usually used when there are no s particular set of classes. It mostly used when the labeling process of data is unsuitable due to have unstructured text. Unsupervised machine learning approaches relay on using clustering algorithms to group data into diverse classes without pre-identifying the label of each class. Cluster analysis approaches can be classified into Hierarchical and Partition clustering. Based on [48] cluster analysis approaches can be classified into Hierarchical and Partition clustering [51] .

2.1) Hierarchical methods

Hierarchical methods produce a hierarchical structure of the data set that is represented by nested groups order as a tree. There are two mains approaches of Hierarchical techniques: Agglomerative and Divisive clustering. Agglomerative approach assumes that each data has its separate cluster then each cluster is merged with other ones based on measuring similarity between related clusters. The process of merging is continuous till all data been in one cluster[52]. The divisive clustering is also can be called the top-down method. This approach starts by grouping all data in one individual cluster. Then these data will spread into sub-clusters over a recursive process by calculate the similarity degree between them [53].

2.2) Partition clustering

Partition clustering is a type of unsupervised learning approach used for sentiment analysis. Its algorithm based on dividing data into separate clusters where each element belongs to one-cluster. The partitioning is performed by calculating the similarity distance between clusters. One of common algorithms for partition technique is the K-means clustering algorithm. In K-means algorithm, it is needed to pre-detect number of initial cluster. After that, each point of data is assigned into different clusters by measuring the nearest distance between the data point and the cluster centroids. This procedure is reduplicated till there is no further change in the centroids. Partition clustering algorithms have been widely used in sentiment analysis in identifying the polarity of product reviews or classifying social media posts based on their sentiment. Other studies used the K-means clustering algorithm to classify textual data such as [54, 55].

Other partitioning algorithms include the Fuzzy C-Means (FCM) clustering algorithm. This algorithm permits to classify data into more than one cluster with a degree of membership. The authors of [56] utilized a fuzzy clustering algorithm to perform sentiment analysis on micro-blogs. Overall, partition methods are useful in situations where the data can be clearly divided into distinct clusters.

B. Lexicon-based approach

Sentiment analysis based on lexicon relies on various attributes and features of textual data, which reflect the emotions, ideas, and meanings expressed in the text. The sentiment analysis based on lexicon uses set of lexical resources such as dictionaries or parts of speech (POS) patterns to assign a specific score for each feature in texts.

The lexical algorithm assumes that the polarity of a text is detected by taking the summation of the sentiment weights of every feature in the text. The final polarity of a text is detected by calculating the values of the semantic orientation of the phrases which it is composed. The text is tokenized into single words and linked it with its sentiment values that taken from the lexicon. Then formula of summation is applied to detect emotion of text [57]. For example, the sentence "she is a bright woman, but she is a boring person" is classified as a negative opinion. This is because the weights of "bright"=2, and "boring"= -3, so the weights of text will be -1. The sentence is categorized as a positive opinion when the weight of the polarity is positive. In other cases, if the weight of the text is equal to zero, the sentence is classified as a neutral emotion. The main issue in this approach is domain dependence. There are set of words have multiple meanings in different sentences

based on purpose such as "the Mobil's cammer is "small" and The TV screen is small". The word "small" in first sentence refers to appositive emotions while the word "small" in second sentence denotes to a negative statues since of most people prefer a wide screen. The two main techniques of lexicon based approach are: Dictionary-based approach and Corpus-based approach. The lexicon-based approach is characterized by being a simple, straightforward and does not need a training phase for data. Therefore it is mostly used as a baseline for sentiment analysis tasks. However, it relies on general preexist dictionaries that does not specialize in a specific field. This make it unable to define specific emotions related to specific vocabulary [18].

1) Dictionary-based approach

The approach assumes that all synonyms phrases carry the same polarities whereas antonyms phrases carry the opposite sentiment polarities. This approach depends on set of dictionaries such as WordNet and SentiWordNet to detect sentiment lexicon of words

In its procedure, Dictionary-based approach finds the pre-known opinion and orientation phrases, and then detects their synonyms and antonyms through lexical resources like dictionaries. Newly incoming words are attracted to the previous list till no new words are detected[58].

Wankhede et al. [59] introduced a well-known dictionary called SentiWordNet 3.0. It is a lexical resource consists from "synsets" which is associated with a positive, negative score start from 1 to 0. The authors in [59] suggested a method to build a thesaurus lexicon by utilized three different online dictionaries. The main obstacles of this approaches is that it may not be able to observe sentiment phrase related to specific domain**.** However, this technique is not expensive as it does not require training dataset [60].

2) Corpus-based approach

Corpus-based approaches launch with a list of opinion phrases and perform syntactic to find other new opinion phrases with their orientation in a huge corpus [61]. Their authors constructed a set of adjectives that occur frequently with their orientation. After that to extend the former set, they considered all words that occurred side by side in the previous patterns have the same orientation. They used a log liner model to determine whether two related adjectives hold the same or different orientation. After that, they performed a clustering algorithm to separate adjectives into two sub cluster of different orientation. Generally it is characterized by simplicity, but it requires a huge data set to determine the polarity of the words and so to discover the polarity of the given text.

A Comparative evaluation of the different learning approach

Machine learning methods such as SVM and Bayes have been effective used for text classification in many studies. Authors in [42] find that the performance of the SVM method with particular features like (unigrams + bigrams) and (adjectives) is the best choice. They record an accuracy rate of 82.7

Abdulla et al [62] have used four classifiers which are: SVM, NB, D-Tree, and KNN. The experiment result shows that the SVM, NB classifiers are more accurate than others. In their study, the accuracy rate of SVM and NB are 84.7 and 80.4 respectively, while D-Tree and KNN are 51.3, 50 respectively

. Zhang et al [63] stated that the Naive Bayes has slightly high effective results within the small corpus such as a short review. Also, they think that SVM is much appropriate when training large corpus. Furthermore, Kalcheva et. [64] have obtained better results with SVM methods rather than Naïve Bayes. Appling supervised and unsupervised learning approaches for text classification in social media based in on set of consideration to achieve better results such as the attributes of the texts being analyzed. Typically, supervised learning gives efficient results with the large textual data [65].

However, this approach depends on train-test conception which could be time-consuming. As menation by[67], the supervised learning approach has the ability to classify unknown documents by using the previous learning rather than the lexicon method which based on deep learning to categorize texts as mentioned in [43] and [68]. On the other hand, unsupervised learning techniques provide deep explanation for the inner structure of patterns within dataset. Furthermore, other researches have shown that unsupervised approaches give better accuracy than supervised methods [62]. According to this study, the main reason for the high efficiency with the lexicon-based approach (unsupervised training) is because of the small size of the lexicon that uses in this study [69]. Typically, Sentiment lexicon is a simple and efficient approach for sentiment analysis that does not require a training stage.

It is possible to manually develop the lexicon approach by associating words with other words in the corpus. Also, it is possible with lexicon analysis to use the saved weight value of sentiment from other dictionaries such as WordNet. However, because of the ambiguity and complexity of web data, we cannot just aggregate the weight of each feature and depend on this result to classify the polarity of texts. This is because that many words could have both negative and positive weight depending on purpose of texts [65]. The authors in[70] and [42] suggested combining each lexicon-based approach and supervised based approach to create a novel hybrid model for sentiment analysis. Their result indicates that this scheme outperformed the supervised approach and lexicon approach when they use it in a separate way.

VI. CONCLUSIONS

In the present study, we highlighted the general research work regarding the text analysis of social media. Many researchers have cleared various models to explain text analysis. In spite of that significant progress in this field has been done with many difficulties that need to be solved.

Sentiment classification has become an active research field that presents unlimited applications with massive challenging problems. This is due to the reality that online information is ambiguous data and needs many filters to prepare data for

classification. This survey deeply discussed the basic requirements to create an integral sentimental analysis system. It also provided a comparison. In this comparison, some studies indicating that the supervised approach mostly produced the best results. For future work, we are interested in examining more classifiers with different features selections such as Bag of Words and POS. Moreover, more potentially complicated cases of texts such as texts contain negation, intensifier, and emotions are aimed to discuss.

References

[1] B. Liu, "Opinion mining and sentiment analysis," in *Sentiment Analysis and Opinion Mining*, vol. 2, Springer Cham, 2018, pp. 413–434. doi: 10.1007/978-3-319-73531-3_13.

[2] V. Ramanathan and T. Meyyappan, "Twitter Text Mining for Sentiment Analysis on People's Feedback about Oman Tourism," *2019 4th MEC International Conference on Big Data and Smart City (ICBDSC)*, Muscat, Oman, pp. 1-5, 2019, doi: 10.1109/ICBDSC.2019.8645596.

[3] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Soc. Netw. Anal. Min.*, vol. 11, no. 1, p. 81, Dec. 2021, doi: 10.1007/s13278-021-00776-6.

[4] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: a review," *Artif. Intell. Rev.*, vol. 53, no. 6, pp. 4335–4385, Aug. 2020, doi: 10.1007/s10462-019-09794-5.

[5] A. Jain, D. Somwanshi, K. Joshi, and S. S. Bhatt, "A Review: Data Mining Classification Techniques," *in 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM)*, Apr. 2022, pp. 636–642. doi: 10.1109/ICIEM54221.2022.9853036.

[6] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, "Sentiment Analysis Based on Deep Learning: A Comparative Study," *Electronics*, vol. 9, no. 3, p. 483, Mar. 2020, doi: 10.3390/electronics9030483.

[7] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing 5Contextual Polarity in Phrase-Level Sentiment Analysis," *Proc. ofHuman Lang. Technol. Conf. Conf. Empir. Methods Nat. Lang.*, no. October, pp. 347–354, 2005.

[8] H. H. Do, P. Prasad, A. Maag, and A. Alsadoon, "Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review," *Expert Syst. Appl.*, vol. 118, pp. 272–299, Mar. 2019, doi: 10.1016/j.eswa.2018.10.003.

[9] K. Cheng, Y. Yue, and Z. Song, "Sentiment Classification Based on Part-of-Speech and Self-Attention Mechanism," *IEEE Access*, vol. 8, pp. 16387–16396, 2020, doi: 10.1109/ACCESS.2020.2967103.

[10] G. Veena, A. Vinayak, and A. J. Nair, "Sentiment Analysis using Improved Vader and Dependency Parsing," in 2021 2nd Global Conference for Advancement in Technology (GCAT), Oct. 2021, pp. 1–6. doi: 10.1109/GCAT52182.2021.9587829.

[11] D. Ekawati and M. L. Khodra, "Aspect-based sentiment analysis for Indonesian restaurant reviews," *Proc. - 2017 Int. Conf. Adv. Informatics Concepts, Theory Appl. ICAICTA 2017*, 2017, doi: 10.1109/ICAICTA.2017.8090963.

[12] Mowlaei, Mohammad Erfan; Abadeh, Mohammad Saniee; Keshavarz, Hamidreza, " Aspect-Based Sentiment Analysis using Adaptive Aspect-Based Lexicons," *Expert Systems with Applications*, vol. 148, 2020, doi:10.1016/j.eswa.2020.113234

[13] C. N. Subalalitha, "Information extraction framework for Kurunthogai," Sādhanā, vol. 44, no. 7, p. 156, Jul. 2019, doi: 10.1007/s12046-019-1140-y.

[14] M. Pantic, N. Sebe, J. F. Cohn, and T. Huang, "Affective multimodal human-computer interaction," in *Proceedings of the 13th ACM International Conference on Multimedia, MM 2005*, 2005, pp. 669–676. doi: 10.1145/1101149.1101299.

[15] X. Farkas and M. Bene, "Images, Politicians, and Social Media: Patterns and Effects of Politicians' Image-Based Political Communication Strategies on Social Media," *Int. J. Press.*, vol. 26, no. 1, pp. 119–142, Jan. 2021, doi: 10.1177/1940161220959553.

[16] F. Falck *et al.*, "Measuring Proximity Between Newspapers and Political Parties : The Sentiment Political Compass," *Policy & Internet*, vol. 10, no. 2, pp. 1–33, 2019, doi: 10.1002/poi3.222.

[17] S. A. Salloum, M. Al-Emran, A. A. Monem, and K. Shaalan, "A survey of text mining in social media: Facebook and Twitter perspectives," *Adv. Sci. Technol. Eng. Syst.*, vol. 2, no. 1, pp. 127–133, 2017, doi: 10.25046/aj020115.

[18] F. Javier, G. Alor-hern, S. Luis, and P. Salas-z, *Use of Sentiment Analysis Techniques in Healthcare Domain*. Springer Nature Switzerland, 2019. doi: 10.1007/978-

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors).*

3-030-06149-4.

[19] N. Alabid and Z. Katheeth, "Sentiment analysis of Twitter posts related to the COVID-19 vaccines," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 24, no. 3, p. 11591, 2021, doi: 10.11591/ijeecs.v24.i3.pp1727-1734.

[20] V. I. . RamyaSri, C. Niharika, and M. Ismail, "Sentiment Analysis of Patients' Opinions in Healthcare using Lexicon-based Method," *Int. J. Eng. Adv. Technol.*, vol. 9, no. 1, 2019.

[21] M. S. K. Abadah, P. Keikhosrokiani, and X. Zhao, "Analytics of Public Reactions to the COVID-19 Vaccine on Twitter Using Sentiment Analysis and Topic Modelling," 2022, pp. 156–188. doi: 10.4018/978-1-6684-5624-8.ch008.

[22] G. L. and Z. H. S. Ji, S. Pan, X. Li, E. Cambria, "Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications," *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 1, pp. 214–226, 2021, doi: 10.1109/TCSS.2020.3021467.

[23] F. Hemmatian and M. Karim, "A survey on classification techniques for opinion mining and sentiment analysis," *Artif. Intell. Rev.*, 2017, doi: 10.1007/s10462-017-9599-6.

[24] J. C. S. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto, "Supervised Learning for Fake News Detection," *IEEE Intell. Syst.*, vol. 34, no. 2, pp. 76–81, Mar. 2019, doi: 10.1109/MIS.2019.2899143.

[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[26] M. Alhawarat and M. Hegazi, "Documents," *IEEE Access*, vol. PP, no. c, p. 1, 2018, doi: 10.1109/ACCESS.2018.2852648.

[27] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Inf. Fusion*, vol. 36, pp. 10–25, 2017, doi: 10.1016/j.inffus.2016.10.004.

[28] M. Parmar, "Sentiment Analysis on Interview Transcripts : An application of NLP for Quantitative Analysis," *2018 Int. Conf. Adv. Comput. Commun. Informatics*, pp. 1063–1068, 2018.

[29] C. Zucco, B. Calabrese, G. Agapito, P. H. Guzzi, and M. Cannataro, "Sentiment analysis for mining texts and social networks data: Methods and tools," *WIREs Data Min. Knowl. Discov.*, vol. 10, no. 1, Jan. 2020, doi: 10.1002/widm.1333.

[30] A. Purpura, and G. Silvello, "Focal elements of neural information retrieval models. An outlook through a reproducibility study," *Inf. Process. Manag.*, vol. 57, no. 6, p. 102109, Nov. 2020, doi: 10.1016/j.ipm.2019.102109.

[31] M. Alassaf and A. M. Qamar, "Improving Sentiment Analysis of Arabic Tweets by One-way ANOVA," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 6, pp. 2849–2859, Jun. 2022, doi: 10.1016/j.jksuci.2020.10.023.

[32] N. Garg and K. Sharma, "Text pre-processing of multilingual for sentiment analysis based on social network data," *Int. J. Electr. Comput. Eng.*, vol. 12, no. 1, p. 776, Feb. 2022, doi: 10.11591/ijece.v12i1.pp776-784.

[33] X. Yue, G. Di, Y. Yu, W. Wang, and H. Shi, P. Berka, "Sentiment analysis using rule-based and case-based reasoning," *J. Intell. Inf. Syst.*, vol. 55, no. 1, pp. 51–66, Aug. 2020, doi: 10.1007/s10844-019-00591-8.

[34] R. Catelli, S. Pelosi, and M. Esposito, "Lexicon-Based vs. Bert-Based Sentiment Analysis: A Comparative Study in Italian," *Electronics*, vol. 11, no. 3, p. 374, Jan. 2022, doi: 10.3390/electronics11030374.

[35] W. Gao, L. Hu, and P. Zhang, "Feature redundancy term variation for mutual information-based feature selection," *Appl. Intell.*, vol. 50, no. 4, pp. 1272–1288, Apr. 2020, doi: 10.1007/s10489-019-01597-z.

[36] Kurniabudi, D. Stiawan, Darmawijoyo, M. Y. Bin Idris, A. M. Bamhdi, and R. Budiarto, "CICIDS-2017 Dataset Feature Analysis With Information Gain for Anomaly Detection," *IEEE Access*, vol. 8, pp. 132911–132921, 2020, doi: 10.1109/ACCESS.2020.3009843.

[37] S. Thaseen and C. A. Kumar, "Intrusion Detection Model Using fusion of Chi-square feature selection and multi class," *J. KING SAUD Univ. - Comput. Inf. Sci.*, vol. 29, no. 4, pp. 462–472, 2017, doi: 10.1016/j.jksuci.2015.12.004.

[38] M. A. Tawhid and A. M. Ibrahim, "Feature selection based on rough set approach, wrapper approach, and binary whale optimization algorithm," *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 3, pp. 573–602, Mar. 2020, doi: 10.1007/s13042-019-00996-5.

[39] H. Liu, M. Zhou, and Q. Liu, "An embedded feature selection method for imbalanced data classification," IEEE/CAA J. Autom. Sin., vol. 6, no. 3, pp. 703–715, May 2019, doi: 10.1109/JAS.2019.1911447

[40] N. Ahmed, J. I. Rafiq, and M. R. Islam, "Enhanced

Human Activity Recognition Based on Smartphone Sensor Data Using Hybrid Feature Selection Model," *Sensors*, vol. 20, no. 1, p. 317, Jan. 2020, doi: 10.3390/s20010317.

[41] T. R. N, "FEATURE SELECTION TECHNIQUES AND ITS IMPORTANCE IN MACHINE LEARNING : A SURVEY," *IEEE Int. Students' Conf. Electr. Electron. Comput. Sci.*, pp. 1–6, 2020, doi: 10.1109/SCEECS48394.2020.189.

[42] B. Pang and L. Lee, "Thumbs up," *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, vol. 31, no. 9. pp. 79–86, 2002. doi: 10.1016/0096-6347(45)90048-2.

[43] P. D. Turney and M. L. Littman, "Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus," *Natl. Res. Counc. Canada*, 2002, doi: 10.4224/8914027.

[44] Sharma, G. Singh, and M. Sharma, "A comprehensive review and analysis of supervised-learning and soft computing techniques for stress diagnosis in humans," *Comput. Biol. Med.*, vol. 134, p. 104450, Jul. 2021, doi: 10.1016/j.compbiomed.2021.104450.

[45] J. Gautam, M. Atrey, N. Malsa, A. Balyan, R. N. Shaw, and A. Ghosh, "Twitter Data Sentiment Analysis Using Naive Bayes Classifier and Generation of Heat Map for Analyzing Intensity Geographically," 2021, pp. 129–139. doi: 10.1007/978-981-33-6919-1_10.

[46] M. Wongkar and A. Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter," in *2019 Fourth International Conference on Informatics and Computing (ICIC)*, Oct. 2019, pp. 1–5. doi: 10.1109/ICIC47613.2019.8985884.

[47] S. S. and P. K.V., "Sentiment analysis of malayalam tweets using machine learning techniques," *ICT Express*, vol. 6, no. 4, pp. 300–305, Dec. 2020, doi: 10.1016/j.icte.2020.04.003.

[48] K. Shankar, S. K. Lakshmanaprabu, D. Gupta, A. Maseleno, and V. H. C. de Albuquerque, "RETRACTED ARTICLE: Optimal feature-based multi-kernel SVM approach for thyroid disease classification," *J. Supercomput.*, vol. 76, no. 2, pp. 1128–1143, Feb. 2020, doi: 10.1007/s11227-018-2469-4.

[49] M. Ahmad and I. Ali, "Sentiment Analysis of Tweets using SVM Sentiment Analysis of Tweets using SVM," no. November, 2017, doi: 10.5120/ijca2017915758.

[50] N. Ruchansky, "CSI : A Hybrid Deep Model for Fake News Detection," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 797–806. doi: 10.1145/3132847.3132877.

[51] N. Alnajran, K. Crockett, D. Mclean, and A. Latham, "Cluster Analysis of Twitter Data : A Review of Algorithms," 2017.

[52] H. Suresh and S. G. Raj, "A Fuzzy Based Hybrid Hierarchical Clustering Model for Twitter Sentiment Analysis," *Springer, Singapore*, vol. 2, pp. 384–397, 2017, doi: 10.1007/978-981-10-6430-2.

[53] M. Yang, Q. Jiang, Y. Shen, Q. Wu, Z. Zhao, and W. Zhou, "Hierarchical human-like strategy for aspect-level sentiment classification with sentiment linguistic knowledge and reinforcement learning," *Neural Networks*, vol. 117, pp. 240–248, Sep. 2019, doi: 10.1016/j.neunet.2019.05.021.

[54] M. W. Nisar, "Opinion mining on large scale data using sentiment analysis and k-means clustering," *Cluster Comput.*, vol. 22, no. 3, 2019, doi: 10.1007/s10586-017-1077-z.

[55] H. Nunoo-mensah, "BIG DATA APPROACH OF SENTIMENT ANALYSIS OF TWITTER DATA USING K- MEAN," *Int. J. Mech. Prod. Eng. Res. Dev.*, vol. 10, no. 3, 2017.

[56] S. Vashishtha and S. Susan, "Fuzzy rule based unsupervised sentiment analysis from social media posts," *Expert Syst. Appl.*, vol. 138, p. 112834, Dec. 2019, doi: 10.1016/j.eswa.2019.112834.

[57] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Syst.*, vol. 226, p. 107134, Aug. 2021, doi: 10.1016/j.knosys.2021.107134.

[58] D. R. Rice and C. Zorn, "Corpus-based dictionaries for sentiment analysis of specialized vocabularies," *Polit. Sci. Res. Methods*, vol. 9, no. 1, pp. 20–35, Jan. 2021, doi: 10.1017/psrm.2019.10

[59] R. Wankhede and A. N. Thakare, "Design approach f or accuracy in movies reviews using sentiment analysis," 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2017, pp. 6-11, doi: 10.1109/ICECA.2017.8203652.

[60] M. Birjali, M. Kasri, and A. Beni-hssane,

"Knowledge-Based Systems A comprehensive survey on sentiment analysis : Approaches , challenges and trends," *Knowledge-Based Syst.*, vol. 226, p. 107134, 2021, doi: 10.1016/j.knosys.2021.107134.

[61] V. Hatzivassiloglou and K. R. Mckeown, "Predicting the Semantic Orientation of Adjectives," *Proc. 35th Annu. Meet. Assoc. Comput. Linguist. Eighth Conf. Eur. Chapter Assoc. Comput. Linguist.*, pp. 174–181, 1997.

[62] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and M. Al-Ayyoub, "Arabic sentiment analysis: Lexicon-based and corpus-based," *2013 IEEE Jordan Conf. Appl. Electr. Eng. Comput. Technol. AEECT 2013*, 2013, doi: 10.1109/AEECT.2013.6716448.

[63] S. Taj, B. B. Shaikh, and A. Fatemah Meghji, "Sentiment Analysis of News Articles: A Lexicon based Approach," in *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, Jan. 2019, pp. 1–5. doi: 10.1109/ICOMET.2019.8673428.

[64] N. Kalcheva, M. Karova, and I. Penev, "Comparison of the accuracy of SVM kemel functions in text classification," in *2020 International Conference on Biomedical Innovations and Applications (BIA)*, Sep. 2020, pp. 141–145. doi: 10.1109/BIA50171.2020.9244278.

[65] C. S. G. Khoo, "Lexicon-based sentiment analysis : Comparative evaluation of six sentiment lexicons," *J. Inf. Sci.*, 2017, doi: 10.1177/0165551517703514.

[67] S. A. S. Neshan and R. Akbari, "A Combination of Machine Learning and Lexicon Based Techniques for Sentiment Analysis," *in 2020 6th International Conference on Web Research (ICWR)*, Apr. 2020, pp. 8–14. doi: 10.1109/ICWR49608.2020.9122298

[68] S. H. Wang, P. Phillips, Z. C. Dong, and Y. D. Zhang, "Intelligent facial emotion recognition based on stationary wavelet entropy and Jaya algorithm," *Neurocomputing*, vol. 272, pp. 668–676, 2018, doi: 10.1016/j.neucom.2017.08.015.

[69] C. C. Aggarwal, "Opinion Mining and Sentiment Analysis," in *Machine Learning for Text*, Cham: Springer International Publishing, 2022, pp. 491–514. doi: 10.1007/978-3-030-96623-2_15.

[70] P. Kumari and M. T. U. Haider, "Sentiment Analysis on Aadhaar for Twitter Data—A Hybrid Classification Approach," 2020, pp. 309–318. doi: 10.1007/978-981-15-0790-8_30.