A Review of Machine Learning Approach for Twitter Sentiment Analysis

Mohammed W. Habib* and Zainab N. Sultani

Computer Science Department, College of Science, Al-Nahrain University, Baghdad-Iraq

Article's Information	Abstract
	One of the active sciences or studies whose importance is rising is the science of sentiment analysis. The reason is due to the increasing sources of data that require investigation. Among the most valuable sources is Twitter, in addition to Facebook and other social media platforms. The objective of sentiment analysis is to classify sentiment/opinions of users as positive, negative, or neutral from textual data. This analysis is valuable for many applications that require understanding people's or users' opinions and emotions about a particular topic, product, or service. Several researchers tackle the problem of sentiment analysis using machine learning algorithms. In this paper, a comparative study is presented of various researches conducted a sentiment analysis on social media and especially on Tweets. The survey carried out in this paper provides an overview of preprocessing steps, machine learning algorithms, and approaches used for sentiment classification during the period 2015-2020.

1. Introduction

Sentiment analysis and opinion mining are the science of studying and analyzing people's opinions, sentiments, and emotions from natural language [1]. It is considered one of the most active and influential research areas in natural language processing (NLP), and it is further considered in data and web mining. Sentiment analysis has expanded to other disciplines such as management and social sciences due to its significant value to the business market and society [2].

The significant development of sentiment analysis coincides with social media's growth, such as reviews, discussions, Twitter, and Facebook, hence having a massive volume of data in digital form for analysis. Analyzing data arises since it is of vital value in: 1) Identifying customer's satisfaction with a specific product/service, 2) Understanding the public opinion about an event or a political issue or topic, 3) even users' views of certain website content [3].

The most distinguished indicators of sentiments are opinion and sentiment words. These words are generally used to represent positive or negative sentiments. For example, *fantastic*, *wonderful*, and *marvelous* are positive, however, *bad*, *poor*, and *disastrous*, are negative sentiment words. Sentiment words are the key to sentiment analysis for obvious reasons. A list of before-mentioned words is called a sentiment/ opinion lexicon [2]. Researchers have designed various algorithms to collect such lexicons. Although sentiment lexicons are essential for sentiment analysis, they are far from sufficient since the sentiment analysis problem is very complicated [2].

Sentiment analysis is usually formulated as a text classification problem that can be solved using machine learning (ML) algorithms. These algorithms usually train sentiment classifiers using unigrams or bigrams that implement various ML methods such as Naive Bayes (NB), Maximum Entropy, or Support Vector Machine (SVM). These ML algorithms are supervised; therefore, they require a labeled dataset. High-quality labeling of a broad training dataset will take time since it is performed manually. The labeled dataset's size and consistency significantly affect the trained model's efficiency [4]. First research in this field was proposed in [2] to classify movie reviews into two categories, positive and negative. It was noted that using unigrams (a bag of words BOW) as features performed quite well with either NB or SVM, even though the authors also tried several other feature options.

One of the most important sources of texts that are read and analyzed is Twitter [5], as it is considered a rich source of people's opinions and thoughts which it is possible to understand, for example, customer satisfaction with a specific product [6].

Twitter facilitates the ease of obtaining data since it provides three types or versions of the application programming interface (API): the representational state transfer (REST) API, the Search API, and the Streaming API. Developers can collect status data and user information through the REST API. The Search API helps developers query unique Twitter content, while the Streaming API can manage real-time Twitter content. Also, developers can combine specific APIs when building their apps; Hence, with the support of massive online data, sentiment analysis appears to have a strong foundation [7].

In this paper, Sentiment Analysis will be discussed, explicitly adopting the machine learning approach. Various researches are considered between the year 2015-2020. This paper focuses on how the sentiment analysis problem was tackled, starting from the preprocessing steps, feature extraction methods, and finally, ML algorithms.

The paper is organized as follows: Section 2 discusses related works, section 3 describes the Machine Learning Approach for Sentiment Analysis. Text Feature Extraction Techniques were discussed in Section 4, while Sentiment Analysis Challenges are discussed in section 5. The experimental results are discussed in Section 6, and then finally, a conclusion is drawn.

2. Related Work

In the past few years, the demand for sentiment analysis has increased and has become a very popular term due to the extensive need to know people's feelings and opinions in many areas therefore recent studies in sentiment analysis has been applied in different languages and for different fields [8] [9], see Table 1. In [10] algorithms were used to deal with text and symbols together instead of relying on one of them because, in most cases, one of them is relied upon. Airline data reviews collected from the Twitter platform were used in this research. In order to extract text features, different techniques such as (TF–IDF, Bag of words, N-gram, and emoticon lexicons) were applied. In [11], a hybrid algorithm was designed ("Hybrid Lexicon-Naive Bayesian Classifier") on data collected from Twitter. The hybrid algorithm provides better results than the two separate methods, where the hybrid method gave 82% while Naive Bayesian and Lexicon had 75% and 61% respectively in terms of accuracy. In [12] For sentiment analysis classification accuracy, we use a hybridization strategy that combines two optimization algorithms and one machine learning classifier, namely particle swarm optimization and genetic algorithm, as well as a decision tree as a classifier. This method is applied to 600 million tweets collected using (URL-based security tool). It was noticed that their method showed an improvement in the accuracy of 86.9%. In [13] authors attempt to uncover, evaluate, and produce recommendations on the thoughts and feelings of people in the texts of their Twitter messages. Where tweets have been gathered for subjects' responses to detect and quantify users' feelings and emotions based on different user-based and Tweet-based parameters. Finally, they used current knowledge to build a general and customized Twitter-based user recommendation system. They have used more than one classifier, but their best results were given by the Naïve Bayes, where the accuracy was 66.86%. Authors in [14] collected data on two halal goods (halal tourism and halal cosmetics) over ten years. An algorithm was used to filter the data using the Twitter search function. Then an experiment is done using deep learning algorithms to quantify and interpret the tweet sensations. Furthermore, convolutional neural networks (CNNs), long-term memory (LSTM), and recurrent neural nets (RNNs) are used to improve predictive accuracy. The findings determine that stacking CNN and LSTM algorithms are correlated with the Word2vec extraction process is the highest precision of 93.78%. A sentiment analysis model in [15] is designed to show which restaurant among McDonald's and Kentucky Fried Chicken (KFC) is more popular. The model incorporates the use of supervised and unsupervised machine learning algorithms. The method of sentiment analysis is as follows: tweet directly derived from the Twitter API, then cleaning and data exploration. After that, for training purposes, the data was fed into multiple models. Each tweet is categorized depending on whether it is a positive, negative, or neutral sentiment. The testing results discovered that for both KFC and McDonald's data, Maximum entropy was the best classifier. In [16] a hybrid approach of sentiment analyzer that includes machine learning has been adopted. They compared sentiment analysis techniques by applying supervised machine learning algorithms such as NB and SVM to analyze political opinions. English and Urdu text data were used, where TF-IDF is implemented to represent the text as feature vectors. In [17] Inflation in social sites made it necessary to pay attention to the topic of sentiment analysis. One of the most important methods of extracting features is the bag of words. They applied NB classifier to 3731 tweets and used TF-IDF as a feature extractor. Sentiment analysis using Doc2Vec on Turkish and English Twitter messages is performed in [18]. Using the semi-controlled learning method, the Doc2Vec algorithm was run on positive, negative, and neutral data, and the findings were recorded. It also demonstrates that the DBOW (Distributed bag of words) technique is more effective than DM (Distributed Memory). The fact that the training and test results obtained for Turkish are lower than those obtained for English indicates that there is a lower causal data set. In [19] authors proposed a Twitter sentiment analysis approach, where rule-based with the poorly supervised Naive Bayes are combined. In the rule-based method a set of rules based on the occurrences are presented to categorize the feelings of tweets, while many emoticons are applied to train the NB classifier. Experiments are conducted on standard emotion 140 datasets and it is represented in BOW feature vector to extract the characteristics. The proposed hybrid method achieve precision equal to 84.96%. Tracking the company's share price relation to customer's opinions is addressed in [20]. To analyze tweets general sentiments, two different textual representations, Word2vec and Ngram are used. They applied sentiment analysis, to analyze the relationship between the stock market movements of a company and the feelings in the tweets. More than one classifier is used, but the random forest was better in terms of accuracy as they got an accuracy equal to 70.49%. They realized that there was a strong relationship between high and low stock prices and general sentiment in the tweets at the end of the study.

Table 1. Comparison of papers from 2015-2020 regarding sentiment analysis on Twitter data.

Paper	Description of datasets	Machine learning techniques	Preprocess	Feature Vector	Languages	Reported accuracy (%)
[10]	14460 reviews	SVM	Remove punctuations, stop words, URL, and digits.	TF-IDF	Various Languages	78%
[11]	1000000 tweets by Twitter streaming API	NB	Remove URL, Retweets, punctuations, stop words, short words, irregular words, and elongated words. Replace Negation words with (not).	TF-IDF	English	75%
[12]	600 million public tweets by Twitter Streaming API	particle swarm optimization + genetic algorithm + decision tree	Remove stop words, expand abbreviation, correct misspelled words, stemming of words, and identification of tags.	TF	English	86.9%
[13]	collection of tweets, comments, retweets, and user information from 25 February 2018 to 8 March 2018	NB	Ignore retweets and remove all user mentions, hashtags, emoticons, multiple occurrences of the same letter, non-alphanumeric characters, and URL.	TF-IDF	English	66.86%
[14]	collected from Twitter using keywords related to halal tourism and halal cosmetics from October 2008 until October 2018	LSTM networks and CNN's.	Eliminate retweets, and the MD5 value of each tweet is computed. Google's algorithm to detect the language of the text	Word2Vec	English and Malay languages	93.78%.
[15]	14000 tweets Collected using Twitter API	maximum entropy	Remove URL, stop words, usernames and accounts, removed numbers and unnecessary spaces, removed punctuations, and convert encoding (Emojis) latin1 to ASCII.	TF-IDF	English	78%
[16]	100,000 tweets	SVM + NB	Translate Urdu to English, Remove URL, mention, special character, stemmer	TF-IDF	English and Urdu	79%
[17]	3731 tweets (Crowdflower's Data for Everyone library)	NB	Remove URL, punctuations, symbols, numbers, Non - English Tweets, stop word, replace emoticons with their sentiment	TF-IDF	English	81.64%
[18]	65,778 tweets Collected from Twitter use API	Semi-supervised	Remove Html Tags, User Name, Hashtags, Phone Number, Area Code, Mail Code, Regular Expression	Doc2Vec	English and Turkish	63.5%
[19]	Stanford sentiment140	combine a rule-based classifier with weakly supervised NB classifier	Lexical Normalization + URL Normalization + Stemming + Stop Word Removal + Special Character Removal	BOW	English	84.96%
[20]	250000 tweets Collected from Twitter use API	Random Forest	Tokenization, Stop words removal, Regex Matching for special character Removal	Word2vec or Ngram	English	70.18%
[21]	Stanford Sentiment Treebank	SVM	Change letters to the Lower case, then remove (URL, stop words, punctuations, username, and repeated letters). In addition to shortcuts such as 20th, 7:35 am, etc.	TF-IDF	English	89.61%
[22]	17,000 tweets collected from various movie tweets	SVM	Remove URL's, Username, Repeated characters, Repeated words, stop words	Word2vec + Ngram	English	84%
[23]	60,000 tweets	SVM	stripped off special characters like '@' and URLs to overcome noise	TF-IDF	English	99%

In [21] the sentiments of the huge tweets generated by the Twitter users stored in the Twitter database are analyzed. The precise methods, NB and SVM, were selected as evaluation criteria for the classification methods. Unigram as well as Bigram have also been used as dimensional extractors along with Chi2 and Singular Value Decomposition. Multiple pre-processing stages, multiple combinations of feature vectors and classification methods are used, 89.61% accuracy was achieved to analyze the sentiment of the Tweets. In [22] sentiment analysis was applied on labeled movie reviews collected from Twitter. Supervised machine-learning algorithms such as SVM, maximum entropy, and NB were implemented to classify data using unigram, bigram, and hybrid features, i.e. unigram + bigram. The result shows that SVM exceeded other classifiers for movie reviews with a remarkable 84% accuracy. Most applications suffer from a lack of training data and therefore they resort to data that does not give the desired results, and this in turn negatively affects the overall accuracy of classifying the text sentiment, so in [23] authors suggested to divide the work into two stages, the first stage is to prepare the data using the Valence Aware Dictionary for Sentiment Reasoning (VADER) for labeling and in the second stage using the labeled tweets as training data for the SVM supervised machine learning algorithm.

3. Discussion and Analysis

As for the analysis of the methods and results mentioned in Table 1, the analysis or comparison is conducted relatively. It is noticed that every research uses different datasets; therefore, the results obtained are based on the dataset used. The analysis of the accuracy results is applied between literatures that uses the same algorithm. For example, [10], [21], [22], [23] used SVM and TF-IDF, however as shown in table 1, authors in [10] obtained less accuracy than the others. This can be for two possible reasons; [10] has fewer preprocessing operations; therefore, unnecessary features remain, and the used dataset could also be the cause of those results. As for research [21], it got better results than research [10] because more preprocessing operations are applied to clean and extract text features. Authors in [22] utilized Word2vec in the process of feature vector representation instead of TFIDF. Word2Vec provided more valuable text features, so it can be concluded that this is one of the reasons this research obtained better results. As for research [23] the high accuracy obtained is due to the used dataset. Researches [11], [13] and [17] used the same technique (NB and TFIDF).

Research [13] obtained the worst accuracy among these three researches. We think this is due to the preprocessing processes used; authors did more preprocessing techniques such as removing all user mentions, hashtags, emoticons and eliminating **multiple occurrences of the same letter, which may cause undefined words.** As for [16], two algorithms (NB and SVM) are used, but the obtained accuracy is not better than other literature for several inevitable reasons. Either because the amount of data is insufficient or that the preprocessing processes (Translate Urdu to English, stemmer) negatively affected the accuracy or combining the two algorithms was not performed effectively.

As for the search [15], the accuracy was 78%, and it is the only research in the table that relied on Maximum entropy and TFIDF. Either the entropy algorithm is unsuitable for the dataset used, or a problem exists with the dataset since it is collected randomly from Twitter using the API.

In the research [20], Random Forest and Word2vec were applied, and the accuracy was (70.18%). Either authors didn't search for the optimal hyperparameters for both algorithms, or the problem was directly from the dataset. In [12], unimaginable accuracy was obtained compared to the rest of the research. The reason for this is attributed either to the innovative method used (particle swarm optimization + genetic algorithm + decision tree) or the nature of the dataset itself.

In [19], very good accuracy was obtained compared to the rest of the research. The reason for this is using a hybrid method, combining a rule-based classifier with a weakly supervised NB classifier.

4. Machine Learning Approach for Sentiment Analysis

There are three approaches in Machine Learning to analyze feelings and opinions, which are: supervised, unsupervised, and semi-supervised learning. Unsupervised learning is used when datasets have no label or class, one of the famous algorithms for unsupervised learning is k-means for clustering problems. As for supervised learning, this technique is used when the datasets have a label or class and the most famous algorithms are Support Vector Machine (SVM), Artificial Neural Network (ANN), Maximum Entropy (ME), Decision Tree (DT), k-Nearest Neighbor (KNN) and Naïve Bayes (NB) classifiers, while in the semi-supervised learning, only part of datasets have the label [24], [25]. In the following sections SVM, NB and KNN are explained briefly:

- Support Vector Machines (SVM): It is a supervised learning model and is considered better than the Naive Bayes in the field of emotional analysis or sentiment analysis as it depends on statistical learning methods [26]. It can be considered a linear or non-linear classifier depending on the kernels. Specifically, it is considered linear when linear Kernel is used as in Figure 1 [27], and nonlinear when (nonlinear) Kernel or Radial Basis Function (RBF) kernel is used [6], [9].
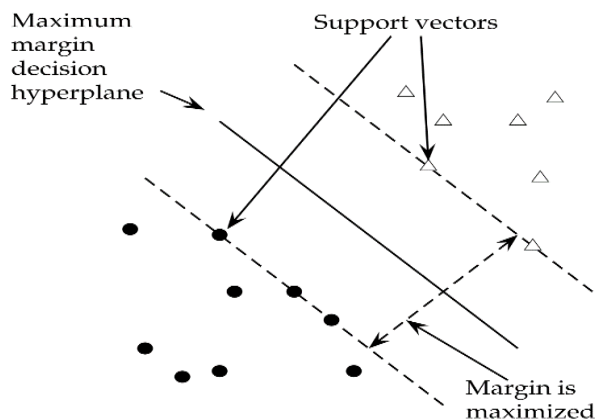
Figure 1. SVM classifier.

- Naive Bayes (NB): It is a simple but effective classifier based on Bayes' theorem, and it depends on the probability in its operation method, see equation 1. NB is simple, fast, accurate, and reliable. It has been successfully used for

many applications, but it works remarkably well with natural language processing (NLP) problems. The word 'Naïve' assumes the features are independent [6], [28]. the following equation is used To calculate the naive bayes:

$$P(\ features\ ) = \frac{P(L)P(L)}{P(\ features\ )} \qquad \ldots(1)$$

where (L|features) denotes the probability of class L given the set of features input (posterior), P(features|L) is the probability of features input given class L (liklihood), while P(features) and P(L) denotes the prior probability of features and class respectively [5].

- K-Nearest Neighbor (KNN): It is one of the models of supervised learning, and the basis of its work relies on the hyperparameter k. The classification of an example (instance) depends on the number of neighbors (k) specified using distance measurements such as Euclidean or Manhattan. The majority of neighbors' class determines the class of the instance, as in Figure 2(a) k is 1 while in Figure 2(b) k is 4. It is necessary to mention that the value of (k) can help eliminate noise points [29], [30].
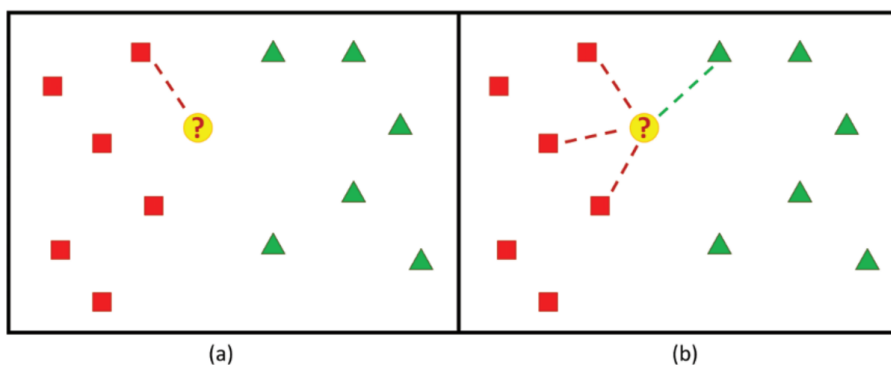

Figure 2. KNN classifier [31].

5. Text Feature Extraction Techniques

It is known that the classifier cannot deal with the text directly but rather deals with numerical values, and from this basis, converting words to feature vectors which they are represented in numeric is a must. The vector representation is one of the pre-processing stages so that the classifier can perform the classification process effectively. There are several techniques used to represent words as numerical values such as Bag of Words (BOW), Term Frequency–Inverse Document Frequency (TF-IDF), and Word to Vector (word2vec) [32].

Bag of Words (BOW): In this method, the text is represented in an array of numbers. The number of rows in this matrix represents the number of texts (tweets, for example), and the number of columns represents the number of essential words (keywords) presented in all texts. Each cell's value can be either 0 or 1, where '0' indicates that a particular keyword doesn't exist in that specific text and '1' if it does. Indeed, with neglecting repetition, one of the biggest problems in this method is the sparsity problem, which means that the matrix has many zeros [33].

Term Frequency–Inverse Document Frequency (TF-IDF): This method is more complex than the (BOW), as it

not only depends on the word's presence in the text but rather depends on the word's repetition in all texts, see equation 2. In this case, the problem of sparsity is reduced, but another problem exists. This problem is the interconnection between the words in the text. For example, if there is a negation before a verb, then the meaning is inverted; however, this is not tackled using this technique. the following equation is used to calculate the TF-IDF:

$$w_{i,j} = tf_{i,j} \times log\ log\ \left(\frac{N}{df_i}\right) \qquad \ldots(2)$$

where (w) denotes the value in which the word will be represented in the matrix, (TF) indicates the number of occurrences of this word in the same text (sentence), (N) represents the number of texts, and (DF) represents the number of occurrences of the word in all texts [34], [35].

Word to Vector (word2vec): In this technique, not all texts (sentences) are represented in an array, but rather every word is represented by a vector. This method appeared to avoid the problem of not representing similar words with similar vectors. There are two training algorithms in word2vec: continuous bag-of-words (CBOW) and skip-gram [36], [37].

6. Sentiment Analysis Challenges

In general, sentiment analysis as a science or research field faces many challenges and difficulties. One of the most critical challenges in sentiment analysis is the accuracy of classifying texts or sentences that are in the form of sarcasm [38], [39]. Besides, it is challenging to obtain good results without the availability of datasets having complicated feelings with the correct label. Usually, such data are few or difficult to provide, so most researchers resort to unsupervised or semi-supervised algorithms since no labels are required in advance. Also, spelling errors and speech in the colloquial language and with a large dimension can cause many problems, such as the difficulty of processing such data [40], [41].

7. Conclusion

Sentiment analysis is significant science. This importance increases with time due to the tremendous expansion in data size, thus the need to analyze and understand it. The difficulties and problems facing the feelings/sentiment analysis arise due to many issues such as ambiguity in interpreting the written text and the curse of dimensionality (features), where computers with higher specification are required. Many techniques can reduce those problems, such as word normalization for text features reduction and implementing recent word embedding techniques to focus on word context. From the researches that we covered in this paper, it can be concluded that the most common supervised learning algorithms used in sentiment analysis are SVM and NB.

References

[1] Yi S. and Liu X.; "Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers' review". Complex Intell. Syst.; 2020, doi: 10.1007/s40747-020-00155-2.

[2] Liu B. and Liu B.; "The Problem of Sentiment Analysis". 2015.

[3] Ramírez-Tinoco F. J.; Alor-Hernández G.; Sánchez-Cervantes J. L.; Olivares-Zepahua B. A. and Rodríguez-Mazahua L.; "A brief review on the use of sentiment analysis approaches in social networks". 2018, doi: 10.1007/978-3-319-69341-5_24.

[4] Dhaoui C.; Webster C. M. and Tan L. P.; "Social media sentiment analysis: lexicon versus machine learning". J. Consum. Mark.; 2017, doi: 10.1108/JCM-03-2017-2141.

[5] Mittal A. and Patidar S.; "Sentiment analysis on twitter data: A survey". 2019, doi: 10.1145/3348445.3348466.

[6] Bhuta S.; Doshi U.; Doshi A. and Narvekar M.; "A review of techniques for sentiment analysis of Twitter data". 2014, doi: 10.1109/ICICICT.2014.6781346.

[7] Fang X. and Zhan J.; "Sentiment analysis using product review data". J. Big Data, 2015, doi: 10.1186/s40537-015-0015-2.

[8] Hercigt T.; Brychcín T.; Svobodat L. and Konkolt M.; "UWB at SemEval-2016 task 5: Aspect based sentiment analysis". 2016, doi: 10.18653/v1/s16-1055.

[9] Gopi A. P.; Jyothi R. N. S.; Narayana V. L. and Sandeep K. S.; "Classification of tweets data based on polarity using improved RBF kernel of SVM". Int. J. Inf. Technol.; 2020, doi: 10.1007/s41870-019-00409-4.

[10] Ullah M. A.; Marium S. M.; Begum S. A. and Dipa N. S.; "An algorithm and method for sentiment analysis using the text and emoticon". ICT Express, 2020, doi: 10.1016/j.icte.2020.07.003.

[11] Rodrigues A. P. and Chiplunkar N. N.; "A new big data approach for topic classification and sentiment analysis of Twitter data". Evol. Intell.; 2019, doi: 10.1007/s12065-019-00236-3.

[12] Nagarajan S. M. and Gandhi U. D.; "Classifying streaming of Twitter data based on sentiment analysis using hybridization". Neural Comput. Appl.; 2019, doi: 10.1007/s00521-018-3476-3.

[13] Sailunaz K. and Alhajj R.; "Emotion and sentiment analysis from Twitter text". J. Comput. Sci.; 2019, doi: 10.1016/j.jocs.2019.05.009.

[14] Feizollah A.; Ainin S.; Anuar N. B.; Abdullah N. A. B. and Hazim M.; "Halal Products on Twitter: Data Extraction and Sentiment Analysis Using Stack of Deep Learning Algorithms". IEEE Access, 2019, doi: 10.1109/ACCESS.2019.2923275.

[15] El Rahman S. A.; Alotaibi F. A. and Alshehri W. A.; "Sentiment Analysis of Twitter Data". 2019, doi: 10.1109/ICCISci.2019.8716464.

[16] Hasan A.; Moin S.; Karim A. and Shamshirband S.; "Machine Learning-Based Sentiment Analysis for Twitter Accounts". Math. Comput. Appl.; 2018, doi: 10.3390/mca23010011.

[17] Malik V. and Kumar A.; "Analysis of Twitter Data Using Deep Learning Approach: LSTM". Int. J. Recent Innov. Trends Comput. Commun.; 2018.

[18] Bilgin M. and Şentürk İ. F.; "Sentiment analysis on Twitter data with semi-supervised Doc2Vec". 2017, doi: 10.1109/UBMK.2017.8093492.

[19] Siddiqua U. A.; Ahsan T. and Chyy A. N.; "Combining a Rule-based Classifier with Weakly Supervised Learning for Twitter Sentiment Analysis". 2017, doi: 10.1109/ICISET.2016.7856499.

[20] Pagolu V. S.; Reddy K. N.; Panda G. and Majhi B.; "Sentiment analysis of Twitter data for predicting stock market movements". 2017, doi: 10.1109/SCOPES.2016.7955659.

[21] Shyamasundar L. B. and Rani P. J.; "Twitter sentiment analysis with different feature extractors and dimensionality reduction using supervised learning algorithms". 2017, doi: 10.1109/INDICON.2016.7839075.

[22] Joshi R. and Tekchandani R.; "Comparative analysis of twitter data using supervised classifiers". 2016, doi: 10.1109/INVENTIVE.2016.7830089.

[23] Ramteke J.; Shah S.; Godhia D. and Shaikh A.; "Election result prediction using Twitter sentiment analysis". 2016, doi: 10.1109/INVENTIVE.2016.

7823280.

[24] Shahnawaz S. and Astya P.; "Sentiment analysis: Approaches and open issues". 2017, doi: 10.1109/CCAA.2017.8229791.

[25] Elamin N.; "Sentiment Analysis with Supervised Learning Techniques: A Survey". Indian J. Sci. Technol.; 2020, doi: 10.17485/ijst/2020/v13i03/148900.

[26] Tyagi P. and Tripathi R. C.; "A Review Towards the Sentiment Analysis Techniques for the Analysis of Twitter Data". SSRN Electron. J.; 2019, doi: 10.2139/ssrn.3368718.

[27] Manning C. D.; Raghavan P.; Schutze H.; Manning C. D.; Raghavan P. and H. Schutze, "Support vector machines and machine learning on documents". Introd. to Inf. Retr.; no. c: 293-320, 2012, doi: 10.1017/cbo9780511809071.016.

[28] Samal B. R.; Behera A. K. and Panda M.; "Performance analysis of supervised machine learning techniques for sentiment analysis". 2017, doi: 10.1109/SSPS.2017.8071579.

[29] Tan S.; "An effective refinement strategy for KNN text classifier". Expert Syst. Appl.; 2006, doi: 10.1016/j.eswa.2005.07.019.

[30] Islam M. J.; Wu Q. M. J.; Ahmadi M. and Sid-Ahmed M. A.; "Investigating the performance of Naive-Bayes classifiers and K- nearest neighbor classifiers". 2007, doi: 10.1109/ICCIT.2007.4420473.

[31] Seidl T.; "Nearest Neighbor Classification". Encycl. Database Syst.; 1: 1885–1890, 2009, doi: 10.1007/978-0-387-39940-9_561.

[32] Zheng A. and Casari A.; "Feature engineering for machine learning". September. 2018.

[33] Kejriwal N.; Kumar S. and Shibata T.; "High performance loop closure detection using bag of word pairs". Rob. Auton. Syst.; 2016, doi: 10.1016/j.robot.2015.12.003.

[34] Dhar A.; Dash N. S. and Roy K.; "Application of TF-IDF feature for categorizing documents of online Bangla web text corpus". 2018, doi: 10.1007/978-981-10-7566-7_6.

[35] Trstenjak B.; Mikac S. and Donko D.; "KNN with TF-IDF based framework for text categorization". 2014, doi: 10.1016/j.proeng.2014.03.129.

[36] M. Bilgin and H. Köktaş, "Sentiment analysis with term weighting and word vectors". Int. Arab J. Inf. Technol.; 2019.

[37] Zhu W.; Zhang W.; Li G. Z.; He C. and Zhang L.; "A study of damp-heat syndrome classification using Word2vec and TF-IDF". 2017, doi: 10.1109/BIBM.2016.7822730.

[38] Khan F. H.; Bashir S. and Qamar U.; "TOM: Twitter opinion mining framework using hybrid classification scheme". Decis. Support Syst.; 2014, doi: 10.1016/j.dss.2013.09.004.

[39] Ghiassi M.; Skinner J. and Zimbra D.; "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network". Expert Syst. Appl.; 2013, doi: 10.1016/j.eswa.2013.05.057.

[40] Zhou S.; Chen Q. and Wang X.; "Active deep learning method for semi-supervised sentiment classification". Neurocomputing, 2013, doi: 10.1016/j.neucom.2013.04.017.

[41] Xiao M. and Guo Y.; "Feature space independent semi-supervised domain adaptation via kernel matching". IEEE Trans. Pattern Anal. Mach. Intell.; 2015, doi: 10.1109/TPAMI.2014.2343216.