How to Cite:
Kaur, C., Boush, M. S. A., Hassen, S. M., Hakami, W. A., Abdalraheem, M. H. O., Galam, N. M., Hadi, N. A., Hadi, N. A., & Benjeed, A. O. S. (2022). Incorporating sentimental analysis into development of a hybrid classification model: A comprehensive study. *International Journal of Health Sciences*, *6*(S1), 1709-1720. https://doi.org/10.53730/ijhs.v6nS1.4924

Incorporating Sentimental Analysis into Development of a Hybrid Classification Model: A Comprehensive Study

Chamandeep Kaur
Lecturer, Computer Science Department, Jazan University, Saudi Arabia

Mawahib Sharafeldin Adam Boush
Assistant Professor, Computer Science Department, Jazan University, Saudi Arabia

Samar Mansoor Hassen
Lecturer, Computer Science Department, Jazan University, Saudi Arabia

Wafaa Abushmlah Hakami
Teaching Assistant, Computer Science Department, Jazan University, Saudi Arabia

Mohammed Hassan Osman Abdalraheem
Assistant Professor, Computer Science Department, Jazan University, Saudi Arabia

Najla Mohammed Galam
Teaching Assistant, Computer Science Department, Jazan University, Saudi Arabia

Nedaa Abdulaziz Hadi
Teaching Assistant, Computer Science Department, Jazan University, Saudi Arabia

Atheer Omar S Benjeed
Lecturer, Computer Science Department, Jazan University, Saudi Arabia

***Abstract*---**The Sentimental Analysis approach is typically used for analyzing a user's ideas, sentiments, and text subjectivity, all of which are expressed through text. Sentimental analysis, also known as "opinion mining," is a type of data mining that follows the concept of emotional analysis presented by people in a thoughtful manner. Based

International Journal of Health Sciences ISSN 2550-6978 E-ISSN 2550-696X © 2022.
***Corresponding author**: Kaur, C.; Email: kaur.chaman83@gmail.com*
Manuscript submitted: 27 Nov 2021, Manuscript revised: 09 Feb 2022, Accepted for publication: 18 March 2022

on historical evidence, websites are the most effective venue for soliciting customer feedback. Existing methodologies based on sentimental analysis are ineffective. As a result, a novel hybrid framework based on three classifiers, including SVM, logistic regression, and random forest, is proposed in this paper. Based on user feedback or historical data, the hybrid model serves as an effective classifier, assisting in the development of more accurate classification results. Furthermore, the proposed model has worked well and has been compared to other methods based on several performance metrics, such as accuracy, precision, recall, and recall.

***Keywords*---**SVM, data analytics, hybrid classification, sentiment analysis, bidirectional long short-term memory, multi-feature fusion, convolution neural network.

Introduction

Information is extracted or mined from large amounts of data using data mining technology, which is a type of technology. This technology makes use of traditional data analytic approaches to find previously unidentified legal incidents and relationships in massive data sets. Statistical models, statistical algorithmic techniques, and machine learning algorithms are some of the most widely used tools. Because of this, in addition to data collection and management, it also performs analysis and forecasting. The goal of this technique is to discover legally permissible, novel, potentially profitable, and understandable correlations, and patterns in existing data [1]. The procedure that is used to identify valuable data patterns is referred to by several distinct names. Initially, mathematicians, database scholars, and business companies began to refer to "data mining" as a term. When it comes to discovering meaningful information from data, the process of "knowledge discovery and discovery" (KDD) encompasses it all. Data mining is an essential part of this entire process. Preparation, selection, and purification of data, as well as an accurate understanding of the findings of the data mining operation, are all important goals of this method. Data mining ensures the discovery and analysis of important information. Data mining is a new method that goes above and beyond what is currently available. Many diverse fields have contributed strategies to data mining.

Individuals' beliefs, behaviors, and attitudes about a specific object or set of circumstances are examined using computers as part of SA. An object can be used to represent people, events, and other subjects. These subjects are commonly discussed by reviewers in their work. Extracting and analyzing other people's thoughts about an item is known as "opinion mining" (OM). On the other hand, SA (Sentiment Analysis) focuses first on determining the emotional content of a text [2]. Aiming to reveal opinions, sentiment analysis aims to identify how strongly people feel and then categorize those feelings accordingly. Document-level sentiment analysis, sentence-level sentiment analysis, and aspect-level sentiment analysis are all terminology used to represent different forms of sentiment analysis. There are three major classification levels for sentiment analysis: That's the major objective of the first strategy, which is to categorize ideas that express either positive or negative thoughts. It sees the information in the document as a whole as being crucial. Sentence-level classification's primary purpose is to perform sentiment categorization across all sentences. This analysis, which is the initial step in the process, determines whether a sentence's subjective or objective nature is with this method, you can identify whether a subjective sentence expresses a good or negative opinion. The primary goal of aspect-level sentiment analysis [3] is to categorize sentiment based on the specific characteristics of an entity. The first stage in this sort of sentiment analysis is to identify the objects and the attributes of the objects that are being considered. When it comes to the characteristics of a similar object, different people may have differing opinions. Machine learning (ML), lexicon-based techniques, and hybrid approaches are the three types of methodologies that can be used to predict attitudes. Implementing the most well-known algorithms while taking advantage of the language's characteristics is the goal of the ML framework. Regarding lexicon in the following section, we will make use of the notion of sentiment lexicon, which is a collection of well-known and previously produced opinionated phrases. Dictionary-based and corpus-based approaches are two types of lexicon-based approaches that can be used in conjunction with one another. These approaches make use of statistical schemes to determine the polarity of sentiments in a specific circumstance. The hybrid approach is a combination of the two techniques. Sentiment lexicons make extensive use of this strategy, which is common. When it comes to practically all tactics, emotional lexicons play a crucial role. Text categorization that makes use of the ML method can be divided into two types [4] on a broad level. The first type of supervised algorithm makes extensive use of tagged training texts in its development. Other alternatives are used in situations where the marked training papers cannot be easily identified by the user. NB is the simplest and most frequently used classification model, as well as one of the most popular. The word distribution is used to compute the posterior class probability in this classifier. To accomplish its objectives, the classifier works in combination with the BOW's feature extraction. There is no distinction between where words exist in the file and where they don't appear in this form of feature extraction. This classifier uses the Bayes Theorem, which is based on the idea of probability distributions, to figure out how likely it is that a certain feature collection is linked to a certain label. By encoding, the Maximum Entropy Classifier converts labelled feature sets to vectors, which are then used to classify the data. Furthermore, this encoded vector is used to compute weights for each trait based on its encoded value. These weights are then combined in order to get the best appropriate label for a feature set. X-weights [5] are used to express this categorization model in a resource. This is used to combine the features of a feature set that has been encoded with an X-encoding. Keep in mind that the encoding oversees converting the C (feature set) pair into a vector. The main goal of the SVM algorithm is to find linear separators in the search space that will separate the different classes in the most efficient way possible while still staying in the search space. Because of the sparseness of the text, SVM is a good fit for text data. Although some qualities are improper, they tend to be associated with them and they are generally sorted out into distinct groups that are divided in a linear fashion. An opinion lexicon is sought for using a lexicon-based strategy. When analyzing text data, the opinion lexicon is used to help guide the researcher. This method is further subdivided into two categories: dictionary-based approaches and corpus-based approaches [6]. The first one seeks out opinionated terms, while the second one consults a dictionary to learn about the synonyms and antonyms of these words, among other things. However, this technique has a significant drawback in that it is unable to recognize opinionated phrases that have a domain and context-based orientation. When combined with the first strategy, the second technique contributes to resolving the difficulty connected with the detection of opinionated words in natural language that have context-specific orientations. When looking for additional opinionated terms in a huge dataset, this approach makes use of the idea of syntactic patterns, which are derived from a fundamental list of opinionated words and can be used to find more opinionated words.

Literature review

Minu Choudhary et al. (2018) compiled evaluations using one of the most popular social media platforms, "Twitter" [7]. 5000 reviews were collected from various mobile phone companies. A Lexicon-based technique was used to conduct sentiment analysis on these ratings. To demonstrate the sentiment analysis findings, a graph was produced. When a consumer was shopping for a new phone, this graph aided in the decision process, and the vendors exploited it to expand their business. Various machine learning algorithms were utilized in the trials to classify the reviews for future study. These machines allowed for a more precise evaluation of the goods.

Rincy Jose et al. (2016) devised an innovative approach for automatically classifying the sentiment of tweets [8]. This strategy was built on machine learning models and a Lexicon. SentiWordNet classifiers, NB classifiers, and HMM classifiers were used in this technique. Following the classification results from these classifiers, the negativity and positivity of each tweet were determined using the majority voting method. These sentiment classifiers were used to mine real-time tweets for political sentiments. This ensemble technique of classifiers contributed in the improvement of sentiment analysis precision. The great accuracy of this method is due to its usage of negation handling and word sense disambiguation.

Vallikannu Ramanathan et al. (2019) suggested a new method of sentiment analysis based on common knowledge [9]. As a result of this technique, Concept Nets were used to create an ontology for Oman's tourism. To begin, the POS was used to categorize the things that were taken from the tweets, which were then analyzed. The concepts in the domain-specific ontology were used to compare these things. A reciprocal sentiment lexicon was used in order to better understand the taken-out creatures' emotions. A final integration was made to bring together explicit attribute and domain semantic inclinations. The conceptual semantic property was used with a machine learning method to help SA do better.

Zahra Rezaei et al. (2017) There were many tweets being posted and transmitted at a rapid rate. The data stream concept has been used to format these messages. Algorithms, in turn, were able to predict Twitter sentiment in real time. The Hoeffding tree algorithm is the most used tool for mining data streams. To pick a splitting feature in a node, this method identified the minimal number of instances that were needed. The Hoeffding tree algorithm did not include MacDiarmid's bound. The McDiarmid tree algorithm was chosen for this reason. Accuracy was comparable to that of the Hoeffding tree for sentiment analysis on Twitter, but processing time was greatly reduced using the McDiarmid tree.

Sonia Saini et al. (2019) advocated for a free-source approach [11]. The twitters were gathered using the API in this method. R programming was used to pre-process, analyze, and visualize these tweets. It was a statistical tool used to analyze the sentiment of tweets. SA was performed using text data extracted from the streamed web. The participants' perceptions were divided into eight distinct groups of sensations. Sentiment analysis was also conducted using two distinct sentiments: positive and negative.

Sahar A. El Rahman et al. (2019) proposed a sentiment analysis algorithm based on real-world Twitter data [12]. Analyzing sentiments from Twitter data was tough because the data was always in an unstructured format. The recommended model, on the other hand, was distinct from past techniques. This model combines supervised and unsupervised approaches. Data collection was conducted on two subjects. McDonald's and Kentucky Fried Chicken were selected as the two most popular eateries. Numerous tests were conducted to evaluate the output of these models. The model presented here was accurate and effective at extracting text from Twitter.

Research Methodology

The methodological framework for the planned method is illustrated in Figure 1, which makes use of both the N-gram and KNN approaches.

Data collection

This study, in particular, provides two distinct sorts of information samples in a physical manner. One of these samples is used for training purposes, while the other is used for testing purposes. The training sample is symmetrical in terms of X and Y. X denotes a feasible estimating comment, whereas Y denotes a positive or negative estimating comment. After obtaining comments from various websites, the testing set is created. The tagging of a comment is done by hand so that we can tell which test samples are negative and which are positive.

Preprocessing of data

Specifically, three distinct preprocessing techniques are applied here: stemming, fault modification, and discontinue statement deletion. The initial method of stemming is to attempt to ascertain an emotional basis. This method eliminates suffixes and a number of related terms. This method has the potential to drastically cut energy and time consumption. Due to the limited use of grammar norms, punctuation, and spelling, it is important to come up with a way to fix mistakes.

Lexical analysis of sentences

Generally, a sentence contains two types of sentiments: positive and negative. Additionally, certain inquiries written by non-emotional users constitute objective phrases. To keep the review size to a minimum, sentences may be separated to keep the overall appraisal size to a minimum. Lexical analysis is the primary function of a compiler. This is accomplished through the use of enhanced source code written as sentences by lexicon preprocessors. The lexical analyzer is in charge of translating a raw byte or a collection of input characters from the source file into a token flow. To do this, the input is segmented and superfluous features are removed. The lexical analyzer generates an error after detecting an error. This provision significantly simplifies work on consecutive syntactical analysis. In the alternative, whitespace and comments can appear everywhere. Lexical analysis is used to classify input tokens into many categories, such as opening brackets, white keywords, and integers. Additionally, the lexical stage provides an advantage by compressing the input size by up to 80%. A lexer may be considered the primary layer of a consistent lexical representation of the input dialectal. Lexical and syntactic analyzers collaborate closely. After reading characters from the source code, the lexical analyzer checks for valid tokens. Then, it transfers data in response to the syntax analyzer's request.

Characteristic extraction

When the characteristics of sample data are extracted, the fundamentals of opinion research become apparent. A word is frequently used to denote an entity feature. The use of POS tagging enables the detection and extraction of all nouns for the purpose of character recognition. In feature extraction, the real feature space is converted to a more compressed novel space. All real attributes are converted to a new compressed space without being deleted. However, the actual traits are obliterated in favor of a more condensed group of representations. This signifies that the input data cannot be processed due to its enormous amount. This is why the data is turned into a new feature set with a reduced number of features. The term "text features" refers to the feature's core component. Feature extraction is a technique for condensing the size of the feature space by selecting a feature group. In feature retrieval, redundant characteristics are eliminated. This procedure improves the accuracy of the learning technique and speeds up its execution. By selecting a document piece, the information can be replicated in the content text. Additionally, weight computation is used to refer to text feature extraction.

Hybrid classifier

To facilitate the classification of data into distinct classes, the hybrid classifier technique is used. The hybrid classifier combines three classifiers: SVM, logistic regression, and random forest. The decision tree approach is extremely popular and is frequently used for classification and prediction. The configuration of the decision tree is identical to that of a tree. Each internal node in this configuration relates to a test on a certain feature. Each branch of this tree represents a test result [13]. Additionally, each leaf node is provided with a class label. This node is referred to as the terminating node. The source data is partitioned into subgroups using a feature value test for tree learning. This procedure is done for each of the resulting subsets. This is known as recursive partitioning. The recursion ends when the subset at a node acquires the same value as the target variable, or when partitioning stops adding value to the predictions. This classifier can be generated without any domain knowledge or parameter settings. As a result, this classifier streamlines the knowledge discovery and analysis process. This classifier can deal with large amounts of data. Generally, this classifier delivers extremely accurate results. This paradigm is often defined by a separating hyperplane. In the presence of labeled training data, this technique constructs an optimal hyperplane. This hyperplane is used to classify new patterns. An SVM model encodes patterns as points in space and maps them in such a way that a clear distinction can be drawn between the patterns of the individual classes. The classifier is predicated on the notion that the purpose of a (natural) class is to provide predictions about the values of characteristics associated with its components. Patterns are classified into classes based on their shared attribute values. These are commonly referred to as natural classes. If an agent is aware of the class, it can make predictions about the values of the other attributes in a Bayesian classifier. Alternatively, Bayes' rule may be used to forecast the class that contains the attribute values. A learning agent is used to generate a probabilistic model, including features in this classification model. This model is used to forecast the classification of an unknown pattern. All classifiers are combined with the voting technique in the final output. When we use the voting approach, we allocate weights to each classifier. The accuracy of sentiment analysis may vary depending on the weights assigned. The distances from the origin of the hyper-planes of the support vectors are:

$$d_+ = \frac{|1-b|}{\|w\|^2}$$

The distance between two planes is:

$$d_- = \frac{|1+b|}{\|w\|^2}$$

Linear regression:

$$Y = b_0 + b_1 \times x_1 + b_2 \times x_2 \ldots\ldots.+b_k \times x_k$$

$$Sigmoid\ function{:}\, p = \frac{1}{1+e^{-Y}}$$

Result written below can be achieved by inserting Y in sigmoid function:

$$\text{In}\left(\frac{P}{1-P}\right) = b_0 + b_1 \times x_1 + b_2 \times x_2 \ldots\ldots.+b_k \times x_k$$

This work uses RF classifier for resolving regression issues. Here, MSE is used to know the branching of data from each node.

$$MSE = \frac{1}{N}\sum_{i=1}^{N} \quad (f_i - y_i)^2$$

Here, N corresponds to the number of points. f_i Is the value returned by the model and y_i is the actual value for data point i

By applying Random Forests based on classification data, the Gini index, or the formula is considered for determining the no, of nodes on a decision tree edge.

$$Gini = 1 - \sum_{i=1}^{C} \quad (p_i)^2$$

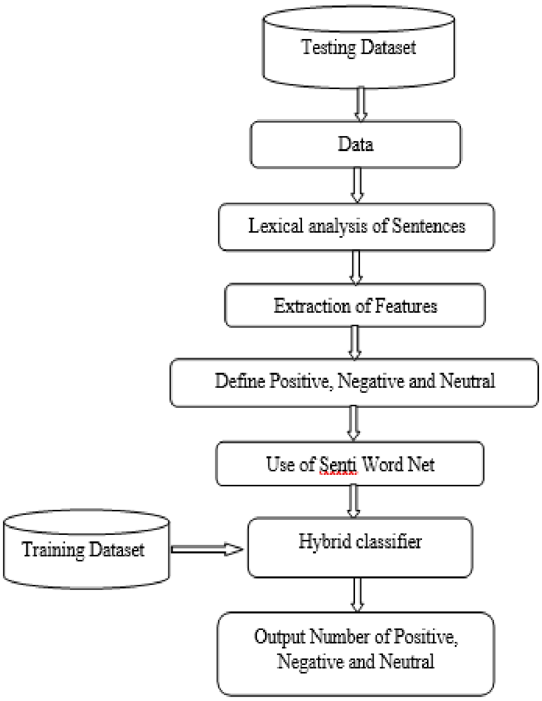
Testing Dataset
Data
Lexical analysis of Sentences
Extraction of Features
Define Positive, Negative and Neutral
Use of Senti Word Net
Training Dataset
Hybrid classifier
Output Number of Positive, Negative and Neutral

Figure 1. Proposed flowchart

Result and Discussion

The country of South Africa is the topic of this book (sentiment analysis). Comparative analysis is carried out between the proposed approach and the SVM approach in this work. To accomplish this, the work makes use of two training and test sets that are diametrically opposed to one another. With respect to execution time and accuracy, the results are compared.

Table 1
Accuracy analysis

Test - Training Ratio	SVM Classifier	Hybrid Classifier
5-95	84%	93%
15-85	83.4%	94.1%
25-75	85%	95%
35-65	83.5%	95%

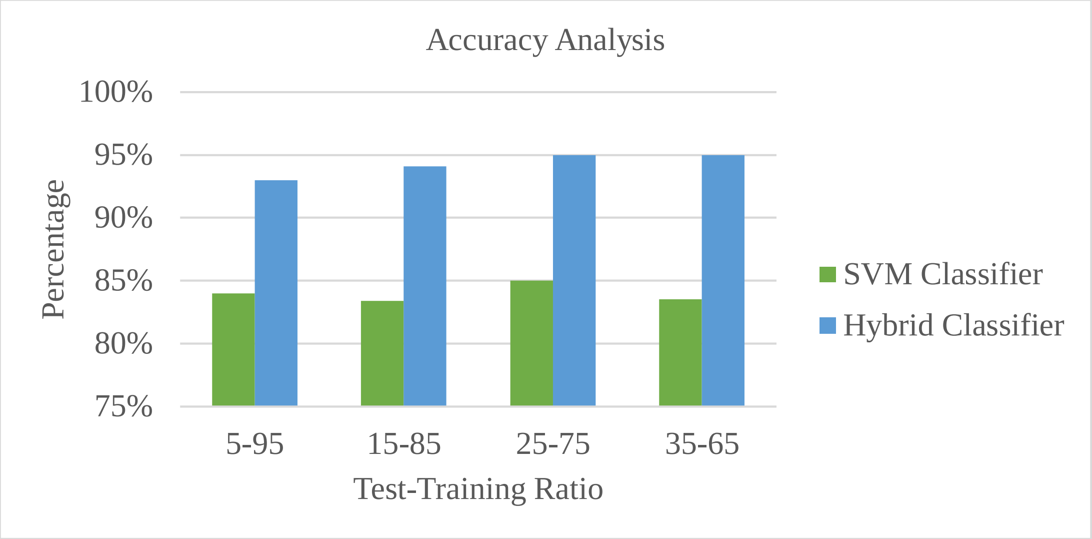

Figure 2. Accuracy analysis

The accuracy-based comparison between the new algorithm and the previous SVM algorithm is depicted in Figure 2. The accuracy of the suggested algorithmic approach for sentiment analysis is higher than the accuracy of the existing algorithmic approach. The outcomes of accuracy testing, and training are compared across multiple sets of training and testing data. The time intervals between the test and training sets are 5:95, 15:85, 25:75, and 35:65. The accuracy achieved by the suggested method on the defined ratios is 93, 94, 95, and 95 %, respectively. When compared to an SVM classifier, the proposed technique produced approximately 10% greater accuracy on average.

Table 2
Precision analysis

Test - Training Ratio	SVM Classifier	Hybrid Classifier
5-95	84%	93%
15-85	83.4%	94.1%
25-75	85%	95%
35-65	83.5%	95%

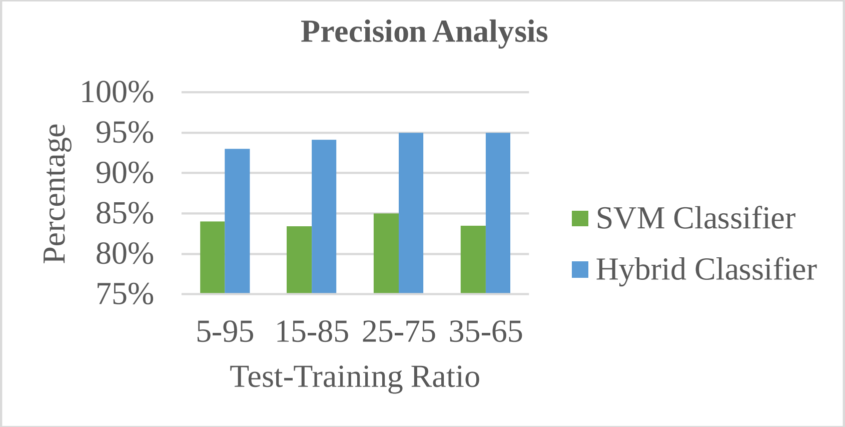

Figure 3. Precision evaluation

Figure 3 depicts a precision-based comparison between the new method and the previous SVM algorithm (precision = 1). The precision of the suggested algorithmic approach for sentiment analysis is higher than the precision of the existing algorithmic approach. The ratios of the test and training sets for sentiment analysis are 5:95, 15:85, 25:75, and 35:65. The precision reached by the suggested method on the defined ratios is 93, 94, 95, and 95 %, respectively. When compared to the SVM classifier, the proposed technique produced around 10 % greater precision.

Table 3
Recall analysis

Test - Training Ratio	SVM Classifier	Hybrid Classifier
5-95	84%	93%
15-85	83.4%	94.1%
25-75	85%	95%
35-65	83.5%	95%

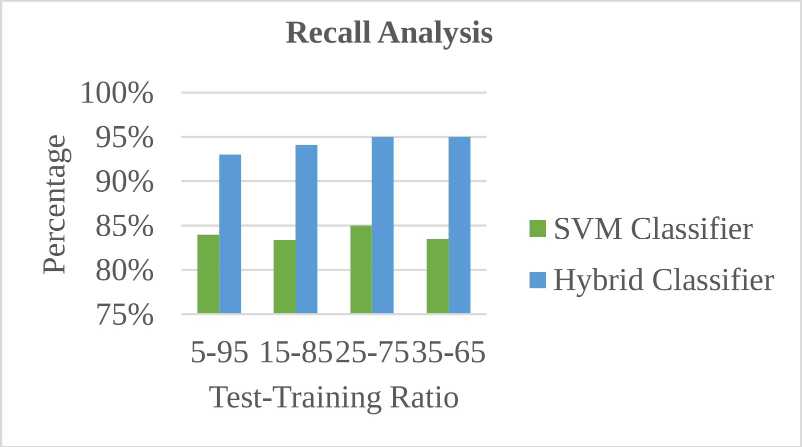

Figure 4. Recall analysis

Figure 4 depicts a recall-based comparison between the new algorithm and the previous SVM algorithm (see text). The recall value of the suggested algorithmic strategy for sentiment analysis is higher than the recall value of the existing algorithmic approach. The time intervals between the test and training sets are 5:95, 15:85, 25:75, and 35:65. The recall achieved by the suggested method on the defined ratios is 93, 94, 95, and 95 %, respectively. The proposed technique outperformed SVM by around 10% in terms of recall rate.

Conclusion

Sentiment analysis is used to analyze the sentiments of input data, as demonstrated in this study. There are several steps in the sentiment analysis process. In this work, a hybrid approach for sentiment analysis is developed. The accuracy and execution of the new technique are evaluated to determine its overall performance. The performance is evaluated using a variety of training and testing ratios. It has been demonstrated that the new technique is more accurate and takes less time to execute than its predecessor.

References

[1] Woldemariam, Y. (2016) 'Sentiment analysis in a cross-media analysis framework'. *IEEE International Conference on Big Data Analysis (ICBDA).*

[2] Fan, X., Li, X., Du, F., Li, X. and Wei, M. (2016) 'Apply word vectors for sentiment analysis of APP reviews'. *3rd International Conference on Systems and Informatics (ICSAI).*

[3] Ding, J., Sun, H., Wang, X. and Liu, X. (2018) 'Entity-Level Sentiment Analysis of Issue Comments'. *IEEE/ACM 3rd International Workshop on Emotion Awareness in Software Engineering (SEmotion).*

[4] Sabra, K.S., Zantout, R.N., El Abed, M.A. and Hamandi, L. (2017) 'Sentiment analysis: Arabic sentiment lexicons'. *Sensors Networks Smart and Emerging Technologies (SENSET).*

[5] Alfarrarjeh, A., Agrawal, S., Kim, S.H. and Shahabi, C. (2017) 'Geo-Spatial Multimedia Sentiment Analysis in Disasters'. *IEEE International Conference on Data Science and Advanced Analytics (DSAA).*

[6] Vanaja, S. and Belwal, M. (2018) 'Aspect-Level Sentiment Analysis on E-Commerce Data'. *International Conference on Inventive Research in Computing Applications (ICIRCA).*

[7] Choudhary, M. and Choudhary, P.K. (2018) 'Sentiment Analysis of Text Reviewing Algorithm using Data Mining'. *International Conference on Smart Systems and Inventive Technology (ICSSIT).*

[8] Jose, R. and Chooralil, V.S. (2016) 'Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble Approach'. *International Conference on Data Mining and Advanced Computing (SAPIENCE).*

[9] Ramanathan, V. and Meyyappan, T. (2019) 'Twitter Text Mining for Sentiment Analysis on People's Feedback about Oman Tourism'. *4th MEC International Conference on Big Data and Smart City (ICBDSC).*

[10] Rezaei, Z. and Jalali, M. (2017) 'Sentiment analysis on Twitter using McDiarmid tree algorithm'. *7th International Conference on Computer and Knowledge Engineering (ICCKE).*

[11] Saini, S., Punhani, R., Bathla, R. and Shukla, V.K. (2019) 'Sentiment Analysis on Twitter Data using R'. *International Conference on Automation, Computational and Technology Management (ICACTM).*

[12] Sahar, A., El Rahman, FeddahAlhumaidiAlOtaibi, Wejdan Abdullah AlShehri. (2019) 'Sentiment Analysis of Twitter Data'. *International Conference on Computer and Information Sciences (ICCIS).*