Nat. Volatiles & Essent. Oils, 2021; 8(5): 1844 - 1854

NVEO
Natural Volatiles & Essential Oils

Hybrid Classification Method for Sentiment Analysis

Pooja Baliyan[1], Pradeep Pant[2], Baldivya Mitra[3], Vimal Kumar[4], Niranjan Lal[5]

[1,2,3,4]Department of Computer Science & Engineering, Meerut Institute of Engineering and Technology, Meerut, UP, India
[5]CSE, School of Engineering and Technology, Mody University of Science and Technology, Lakshmangarh, Sikar, Rajasthan, India
[1]poojachaudhary500@gmail.com, [3]baldivya.mitra@miet.ac.in, [4]vimal.kumar@miet.ac.in,
[,5]niranjan_verma51@yahoo.com*

ABSTRACT:

Sentiment Analysis is a computer-based approach of analysing opinions, sentiments and text subjectivity. Opinion Mining is the other name given to this approach. This approach follows the notion of emotion analysis expressed by peoples in careful way. In order to get opinionated data, websites are the best place. The proposed technique has various steps. This work is based on hybrid classification method for Sentiment Analysis. The performance of devised technique is examined with respect to some important metrics. The devised method performs better with regards to both of these metrics.

KEYWORDS: SVM, Hybrid Classification, Sentiment analysis

1. INTRODUCTION

Data mining as a technology extracts or mine information from great quantity of data. This technology applies classy data analytic methods for discovering the earlier unidentified, lawful instances and associations in immense records. Some popular tools are statistical models, statistical algorithmic approaches, and ML algorithms. Hence, data mining not only collects and manages data, but also do analysis and forecasting. This technology aims to detect legal, new, possibly valuable, and comprehensible associations and patterns in available data [1]. The process applied to find valuable data patterns is identified by different names. At first, mathematicians, database scholars, and the commercial organizations started to use the word “data mining”. The word KDD represents the whole method of exploring valuable knowledge from data. In this whole process, data mining is a major step. The main tasks in this method are preparation, selection and cleaning of data with correct understanding of the outcomes of the data mining method. Data mining ensures the determining of useful knowledge. Data mining is an advancement of conventional schemes. The methods derived from several disciplines are included in data mining.

SA refers to the computer-based analysis of individual’s beliefs, behaviours and sentiments with respect to an object. The object can denote people, incidents or subjects. These subjects are generally enclosed in reviews. OM (Opinion Mining) does the extraction and analysis of people’s opinion regarding an object. On the other hand, SA (Sentiment Analysis) first recognizes the sentiment articulated in a text then analyze it [2]. Hence, sentiment analysis aims to discover opinions, recognizes the sentiments expressed by people, and then classifies the sentiments based on polarity [13]. Sentiment analysis includes three major classification levels. These are termed as document-level, sentence-level, and aspect-level sentiment analysis. The main objective of first approach is the classification of those opinions that express two types of sentiments either positive or negative. It regards the entire document as fundamental information part. The main focus of sentence-level is to

perform classification of sentiments articulated in all sentences. This analysis firstly gives information about the subjective or objective nature of a sentence. This approach determines the positive or negative opinion expressed in a subjective sentence. Classifying the sentiment based on the particular traits of an entity is the key objective of aspect-level sentiment analysis [3]. Identifying the objects and their aspects, is the primary step in this type of sentiment analysis. People may have dissimilar views with regard to the features of the similar object. The approaches for classifying sentiments are generally categorized into ML, lexicon-based and hybrid approaches. ML scheme implements the eminent algorithms and makes use of language characteristics. The next approach of lexicon is based on a sentiment lexicon which represents the gathering of recognized and already compiled opinionated terms. Lexicon-based Approach can be further partitioned into dictionary and corpus-based approaches. These approaches make usage of statistical schemes for finding the polarity of sentiments. The hybrid Approach integrates both approaches [16]. The use of this approach is quite popular with sentiment lexicons. Sentiment lexicons contribute significantly in almost all techniques. The text classification that makes use of ML algorithm can be generally categorized into two categories [4]. The first category of supervised algorithms uses ample amount of labelled training documents. The other approaches are employed in a situation when these labelled training documents cannot be detected easily. NB is the humblest and one of the most popular classification models. This classifier measures the posterior class probability according to the allocation of words in a file. The classifier performs with the BOWs feature extraction. This type of feature extraction does not care about the location of words in the file. This classifier follows the concept of Bayes Theorem for predicting the probability that a specified feature set is related to a specific label. Maximum Entropy Classifier does the conversion of labelled feature sets to vectors by encoding. Further, this encoded vector is adopted for computing weights for every trait. These weights be integrated for determining the most possible label for a feature set. This classification model is represented by a resource of X{weights} [5]. This is employed for combining the features of a feature-set using an X{encoding}. Particularly, the encoding does the mapping of C {(feature set, label)} pair to a vector. Deriving linear separators in the search space that separate the several classes in optimal manner is the main objective of SVM classifier. Text data are preferably suitable for SVM due to the sparse text, where some traits are inappropriate, but they tend to be associate and normally sorted out into different groups separated in linear manner. Lexicon-based approach aims to find the opinion lexicon. The opinion lexicon is employed for the analysis of text data. This approach is further divided into dictionary and Corpus based approaches [6]. The first one aims to find the opinionated words, and then looks into the dictionary for exploring the synonyms and antonyms of these words. This approach has a main drawback that it cannot find opinionated words with domain and context-based orientation. The second approach contributes to resolve the issue related to the discovery of opinionated words with context specific orientations. This approach uses the idea of syntactic patterns that ensue collectively by a basic list of opinionated words for finding other opinionated words in a big data.

2. LITERATURE REVIEW

Minu Choudhary, et.al (2018) extracted reviews from "Twitter", that was one of the widespread social media platforms [7,14]. There were 5000 reviews collected from various brands of mobiles. The Lexicon based technique was utilized for the sentiment analysis of these reviews. A graph was prepared to represent the result of sentiment analysis. To buy any new mobile phone, this graph helped the purchaser in decision making and the sellers had used it to enhance their business. For the future analysis, the diverse machine learning methods were utilized in the experiments for the classification of the reviews. The more accurate opinion regarding the product had been acquired from these machines.

Rincy Jose, et.al (2016) devised a new scheme for robotically classifying the sentiment of tweets [8]. In this approach, ML models and Lexicon-based approach were utilized together. SentiWordNet classifier, NB classifiers and HMM classifiers had applied in this technique. After achieving the results of reviews from these classifiers, the negativity and positivity of each tweet was determined on the basis of majority voting principle. The political sentiments had been found from the real time tweets by utilizing these sentiment classifiers. This ensemble technique of classifiers helped to acquire an enhanced exactness in sentiment analysis. High accuracy had been achieved by using negation handling and word sense disambiguation of this method.

Vallikannu Ramanathan, et.al (2019) suggested a novel technique for sentiment analysis that was based on the common knowledge [9]. A Concept Net based Oman tourism ontology had been produced in this technique. Firstly, the POS was utilized to identify the entities that were taken from the tweets. In the domain specific ontology, these entities were compared with the concept. The sentiments of the taken-out entities were further compared by utilizing mutual sentiment lexicon approach. At last, semantic orientations of explicit attributes and domain were combined. To improve the performance of SA, the conceptual semantic as an attribute had combined with the machine learning algorithm.

Zahra Rezaei, et.al (2017) studied that the messages on the Twitter were continually generated and they were reached to the destination at high speed [10]. These messages had followed the data stream model. Therefore, the algorithms were utilized to foresee sentiment on Twitter under limited and real time[15]. For this, the most well-liked tool in mining data streams that was known as Hoeffding tree algorithm was utilized. By this algorithm, the smallest number of instances had been found that were required to select a splitting feature in a node. In the Hoeffding tree algorithm, MacDiarmid's bound was replaced. That was why, McDiarmid tree algorithm had been applied in this paper. The accuracy acquired on Twitter for sentiment analysis from the McDiarmid tree was similar to that of Hoeffding tree and the processing time of the former was reduced noticeably.

Sonia Saini, et.al (2019) recommended an open source technique [11]. In this approach, the tweets were collected from the Twitter API. These tweets were pre-processed, analysed and visualized by utilizing R programming. It was a statistical tool that was applied for the sentimental analysis of tweets. SA was done on the basis of the text data that was retrieve from the streamed web. The perceptions of the people were classified according to the eight different categories of the feelings. The analysis of sentiments was also based on the two unique sentiments: positive and negative.

Sahar A. El Rahman, et.al (2019) presented a model in which sentiment analysis was based on the real data that was collected from Twitter [12]. It was hard to analyse the sentiments from the data collected from twitter as this data was always present in unstructured form. But the recommended model was different from former methods. In this model, supervised and unsupervised algorithms were combined together. The data was gathered on two subjects. Two restaurants McDonalds and KFC were selected to represent which one was more popular. Different testing metrics were utilized for inspection of the result from these models. The presented model performed accurately and successfully on mining texts that were directly extracted from Twitter.

3. RESEARCH METHODOLOGY

The methodical system for projected method is elucidated in the figure 1 where both N-gram and KNN approaches are exploited.

A. Dataset

In particular, this work creates two types of information samples in physical manner. One of these samples is employed for training while other one performs testing. The training sample has X: Y relation. X signifies feasible estimating comment while Y estimates positive or negative comment. Once the comments from different websites are obtained, the testing set is created. The tagging of a comment is carried out in manual way so that the negative or positive test samples can be recognized.

B. Data Pre-processing

Specifically, three dissimilar pre-processing approaches of Stemming, fault alteration and discontinue statement elimination are used here. The first approach of stemming aims to figure out a sentiment base. This approach eradicates suffixes and several related terms. This approach can significantly reduce the amount of energy and time consumed. It is essential to develop fault alteration approach due to the limited use of grammatical rules, punctuation and spellings.

C. Lexical Analysis of Sentences

A sentence generally contains two types of sentiments either positive or negative. Also, some queries written by users without any emotions represent objective sentences. To make the review size minimal, sentences may be separate for the minimization of overall appraisal size. The primary step of a compiler is lexical analysis. Improved source code written as sentences by lexicon pre-processors is used for this purpose. Lexical analyser is responsible for converting a rough byte or a set of input characters approaching from the origin file into a token flow. In order to do so, the input in divided into various parts and redundant features are eliminated. Lexical analyser causes error after finding an illegal token. This provision simplifies works for the consecutive syntactical analysis to a significant level. In other case, whitespace and comments can take place everywhere. Lexical analysis aims for classifying input tokens into different types including opening bracket, white keyword, integer etc. lexical stage offers one more advantage by compressing input size up to 80%. The prime layer of a consistent lexical view on the input dialectal might be considered by a lexer. The lexical and syntax

analysers work closely. Lexical analyser after reading characters from the source code, verifies for valid tokens. Then, it transfers data when demanded by syntax analyser.

D. Extraction of Features

The core concerning opinion study raise when features of sample data are extracted. A noun is utilized all the time for representing an entity feature. The utilization of POS tagging is performed for the detection and extraction of all the nouns for the recognition of all characteristics. The conversion of real feature space is carried out into a more compressed novel space in feature extraction. The conversion of all real attributes is carried out into new compressed space with no deletion. The real features, however, are replaced by a shorter group of representation. This means that it is not possible to process the input data because of huge volume. This is the reason that this data is converted into a new feature set containing less no. of features. Text features refer to the fundamental element of the feature. Feature extraction represents a process applied for reducing the size of feature space by choosing a feature group. The eradication of redundant features is carried out in feature retrieval. This process increases the accuracy rate of learning approach and reduces execution time. The information can be imitated on the content text by selecting document portion. Also, text feature extraction is referred as weight computation.

E. Hybrid Classifier

In order to facilitate classification of data into certain classes, technique of hybrid classifier is applied. The hybrid classifier is the combination of three classifier which are SVM, logistic regression and Random forest. Decision tree is a very popular approach which is generally used for classification and prediction. The configuration of Decision tree is just like a tree. In this configuration, each internal node refers to a test on a feature. All branches of this tree represent the tested outcomes [13]. Also, a class label is assigned to every leaf node. This node is known as terminal node. The partitioning of source set is performed into subsets on the basis of a feature value test for tree learning. This process is repeated again and again on all resultant subsets. This process is referred as recursive partitioning. The recursion is finished after the subset at a node get the value same as the target variable, or when partitioning no longer inserts value to the predictions. There is no need of any domain knowledge or parameter setting for generating this classifier. Therefore, this classifier simplifies the analysis process of knowledge discovery. This classifier is able to handle the huge volume of data. This classifier generally produces very accurate outcomes. A separating hyperplane generally defines this model. This algorithm generates an optimum hyperplane in the presence of labelled training data. New patterns are categorized using this hyperplane. An SVM model represents patterns as points in space and maps them so that a clear gap can divide the patterns of the individual classes in the widest way. The classifier depends on the concept that the function of a (natural) class is to generate prediction about the values of features for component of that class. The grouping of patterns is done into classes as they contain similar values for the attributes. These classes are generally referred as natural classes. In a Bayesian classifier if an agent knows the class, it can make prediction about the values of the other attributes. In other case, Bayes' rule may be applied for predicting the class providing the attribute values. In this classification model, a probabilistic model containing features is constructed using a

learning agent. This model is employed for predicting the classification of a novel pattern. In the last output of all the classifiers get merged with the voting method. When we are applying voting method, weights are assigned to each classifier. According to assigned weights accuracy may varies to analyse sentiments.

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:

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

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

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

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 on the basis of classification data, the Gini index, or the formula is considered for determining the no, of nodes on a decision tree edge.

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

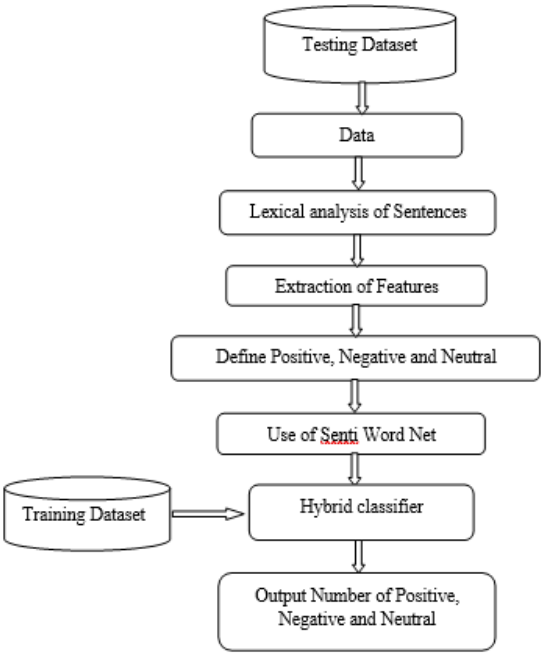

Figure 1: Proposed Flowchart

4. RESULT AND DISCUSSION

The focus of this work is on SA (sentiment analysis). This work performs comparison between the proposed and SVM approach. In order to do so, thus work makes use of two dissimilar training and test sets. The outcomes are compared with regard to execution time and accuracy.

Table 1: Accuracy Analysis

Test-Training Ratio	SVM Classifier	Hybrid Classifier
10-90	83 percent	92 percent
20-80	82.5 percent	93.2 percent
30-70	84 percent	94 percent
40-60	82.6 percent	94. percent

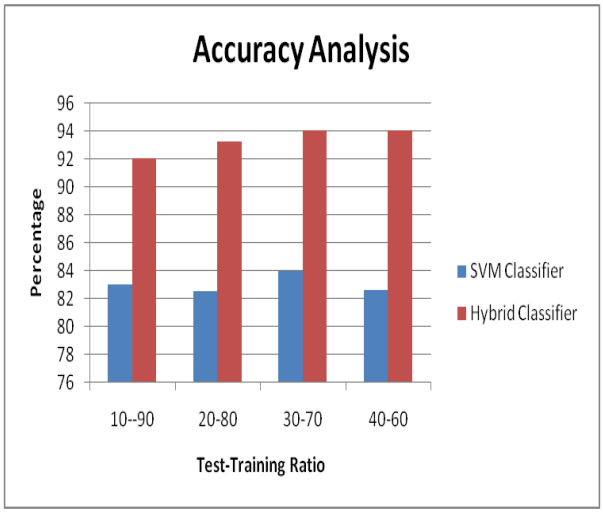

Figure 2: Accuracy Analysis

Figure 2 shows accuracy-based comparison between new algorithm and former SVM algorithm. The accuracy of proposed algorithmic approach is higher than existing algorithmic approach for sentiment analysis. The results of accuracy are analysed on different sets of training and testing. The test and training sets are 10:90, 20:80,30:70 and 40:60. The accuracy which is achieved on defined ratios is 92,93,94,94 by proposed method respectively. The proposed method achieved approx 10 percent more accuracy as compared to SVM classifier.

Table 2: Precision Analysis

Test-Training Ratio	SVM Classifier	Hybrid Classifier
10-90	83.2 percent	92.1 percent
20-80	82.3 percent	93.1 percent
30-70	84.1 percent	94 percent
40-60	82.6 percent	94. percent

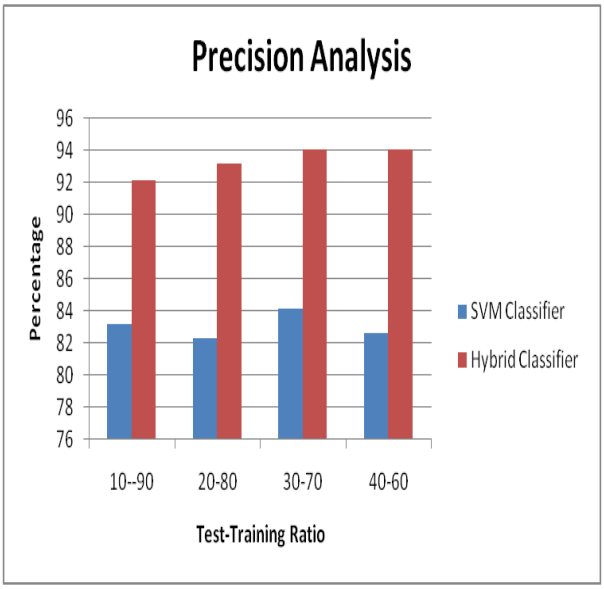

Figure 3: Precision Analysis

Figure 3 shows precision-based comparison between new algorithm and former SVM algorithm. The precision of proposed algorithmic approach is higher than existing algorithmic approach for sentiment analysis. for sentiment analysis, the test and training sets are 10:90, 20:80,30:70 and 40:60. The precision which is achieved on defined ratios is 92,93,94,94 by proposed method respectively. The proposed method achieved approx 10 percent more precision as compared to SVM classifier.

Table 3: Recall Analysis

Test-Training Ratio	SVM Classifier	Hybrid Classifier
10-90	83.2 percent	92.1 percent
20-80	82.3 percent	93.1 percent
30-70	84.1 percent	94 percent
40-60	82.6 percent	94. percent

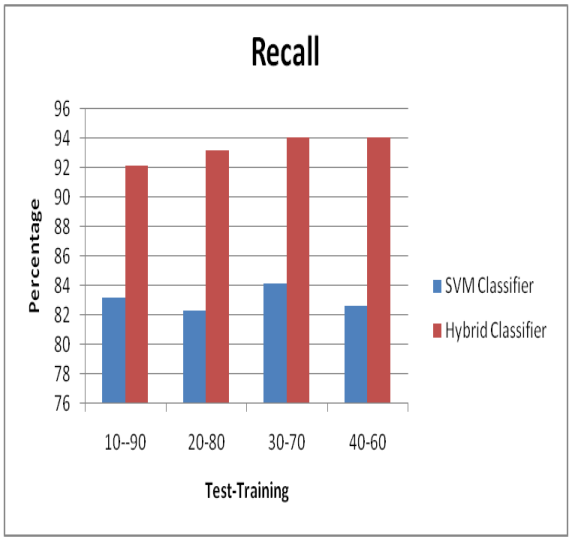

Figure 4: Recall Analysis

Figure 4 shows recall-based comparison between new algorithm and former SVM algorithm. The recall value of proposed algorithmic approach is higher than existing algorithmic approach for sentiment analysis. The test and training sets are 10:90, 20:80,30:70 and 40:60. The recall which is achieved on defined ratios is 92,93,94,94 by proposed method respectively. The proposed method achieved approx 10 percent more recall than SVM.

CONCLUSION

This work concludes that sentiment analysis is used to analyse sentiments of input data. The sentiment analysis has various steps. This work devises a hybrid technique of sentiment analysis. The performance of new technique is analysed in terms of accuracy and execution. The performance is analysed on different training and test ratios. The new technique shows more accuracy and lesser execution time than its counterpart.

References

[1]. Yonas Woldemariam, “Sentiment analysis in a cross-media analysis framework”, 2016, IEEE International Conference on Big Data Analysis (ICBDA)
[2]. Xian Fan, Xiaoge Li, Feihong Du, Xin Li, Mian Wei, “Apply word vectors for sentiment analysis of APP reviews”, 2016, 3rd International Conference on Systems and Informatics (ICSAI)
[3]. Jin Ding, Hailong Sun, Xu Wang, Xudong Liu, “Entity-Level Sentiment Analysis of Issue Comments”, 2018, IEEE/ACM 3rd International Workshop on Emotion Awareness in Software Engineering (SEmotion)
[4]. Khaled S. Sabra, Rached N. Zantout, Mohamad A. El Abed, Lama Hamandi, “Sentiment analysis: Arabic sentiment lexicons”, 2017, Sensors Networks Smart and Emerging Technologies (SENSET)
[5]. Abdullah Alfarrarjeh, Sumeet Agrawal, Seon Ho Kim, Cyrus Shahabi, “Geo-Spatial Multimedia Sentiment Analysis in Disasters”, 2017, IEEE International Conference on Data Science and Advanced Analytics (DSAA)

[6]. Satuluri Vanaja, Meena Belwal, “Aspect-Level Sentiment Analysis on E-Commerce Data”, 2018, International Conference on Inventive Research in Computing Applications (ICIRCA)

[7]. Minu Choudhary, Prashant Kumar Choudhary, “Sentiment Analysis of Text Reviewing Algorithm using Data Mining”, 2018, International Conference on Smart Systems and Inventive Technology (ICSSIT)

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

[9]. Vallikannu Ramanathan, 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)

[10]. Zahra Rezaei, Mehrdad Jalali, “Sentiment analysis on Twitter using McDiarmid tree algorithm”, 2017, 7th International Conference on Computer and Knowledge Engineering (ICCKE)

[11]. Sonia Saini, RituPunhani, Ruchika Bathla, Vinod Kumar Shukla, “Sentiment Analysis on Twitter Data using R”, 2019, International Conference on Automation, Computational and Technology Management (ICACTM)

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

[13]. N. Lal, H. Garg, "Data Analysis: Opinion Mining and Sentiment Analysis of Opinionated Unstructured Data" Springer Nature Singapore Pte Ltd. 2018, Communications in Computer and Information Science (CCIS) Volume 906, pp. 249–258, Nov, 2018.

[14]. N. Lal, N, Kaur “Clustering of Social Networking Data using SparkR in Big Data" Springer Nature Singapore Pte Ltd. 2018, Communications in Computer and Information Science (CCIS), Volume 906, pp. 217–226, Nov, 2018.

[15]. Gupta, Itisha & Joshi, Nisheeth. (2019). Enhanced Twitter Sentiment Analysis Using Hybrid Approach and by Accounting Local Contextual Semantic. Journal of Intelligent Systems. 29. 10.1515/jisys-2019-0106.

[16]. Kumari P., Haider M.T.U. (2020) Sentiment Analysis on Aadhaar for Twitter Data—A Hybrid Classification Approach. In: Bhalla S., Kwan P., Bedekar M., Phalnikar R., Sirsikar S. (eds) Proceeding of International Conference on Computational Science and Applications. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-15-0790-8_30