JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 04, 2020

An efficient Decision Tree Algorithm for analyzing the Twitter Sentiment Analysis

[1]S.Kasthuri, [2]Dr.A.Nisha Jebaseeli,

1Research Scholar, Department of Computer Science, Govt. Arts & Science College, Lalgudi, Trichy, Tamilnadu, India. vishnugka@yahoo.co.in
2Assistant Professor & Head, Department of Computer Science, Govt. Arts & Science College, Lalgudi, Trichy, Tamilnadu, India. nishamarcia@gmail.com

Received: 14 Feb 2020 Revised and Accepted: 25 March 2020

ABSTRACT: Opinion mining and sentiment analysis are valuable to extract the useful subjective information out of text documents. The huge amount of information from this medium has become an attractive resource for organizations to monitor the opinions of users, and therefore, it is receiving a lot of attention in the field of sentiment analysis. However, performing sentiment analysis is a challenging task for the researchers in order to find the users sentiments from the large datasets, because of its unstructured nature, slangs, misspells and abbreviations. To address this problem, a new proposed system is developed in this research study. Here, the proposed system comprises of four major phases; data collection, pre-processing, key word extraction, and classification. Initially, the input data were collected from the twitter dataset. After collecting the data, pre-processing was carried-out for enhancing the quality of collected data. The pre-processing phase comprises of two systems; lemmatization, and removal of stop-words and URLs. Then, an effective topic modelling approach Latent Dirichlet Allocation (LDA) was applied to extract the keywords and also helps in identifying the concerned topics. The extracted key-words were classified into three forms (positive, negative and neutral) by applying an effective machine learning classifier: Decision Tree (DT). The experimental outcome showed that the proposed system enhanced the accuracy in sentiment analysis up to 6-20% related to the existing systems.

KEYWORDS: *Sentiment Analysis, Decision Tree, Twitter, Lemmatization, Latent Dirichlet Allocation*.

I. INTRODUCTION

Twitter is the vital platform used for communication and sharing information with friends [1]. Twitter makes the information easy to spread and read, because it allows the user to publish only 140 characters in a single tweet [2]. Several applications such as elections, reviews, sentiment analysis (SA) and marketing used Twitter as a blogging platform, because it provides a vast amount of data [3-4]. Generally, people tweets about various topics such as reviews on movies, products, brands, politics, etc on social media to share their opinions. [5]. This research work is mainly used to identify the public opinion on various topics in India. It is necessary to determine if the sentences are subjective or objective before the polarity of the sentences can be analysed. Only the subjective sentences will then be further analysed to determine whether they are positive, negative or neutral. Recent methodologies in Twitter Sentiment Analysis (TSA) extracts the twitter text from online blogs, for classifying the text as positive or negative [6-7]. Challenges faced by the researchers in TSA are: neutral tweets are more common than positive and negative ones, which is very difficult to classify. Tweets are very short and often show limited sentiment cues [8].

Many researchers focused on the use of traditional classifiers, like naive Bayes, maximum entropy, and support vector machine (SVM) to solve these problems [9-10].In order to improve the classification accuracy, a new classification methodology is implemented. In this experimental research, TSA is performed on the simulated Sanders twitter data. The proposed methodology consists of two phases such as, pre-processing and classification. The unwanted noises such as URLs, positive and negative emoji, stop-words are reduced by processing the twitter data which was considered as a first phase. The Decision tree (DT) is used for TSA on pre-processed twitter data. The sparse categorical cross-entropy is used to predict the loss of data, whereas existing methods used the binary methods which produced some loss of data and errors in classification. These twitter words are stored in the dictionary with an individual weight value, then the testing data is matched with the dictionary in order to evaluate the performance of the proposed DT approach. In the experimental analysis, the analysis of Decision tree are validated against Random Forest (RF) and Multi-layer Perceptron (MLP).

The rest of the paper is organized as follows: Section 2 describes the survey of recent techniques to classify the tweets. The problem statement is described in section 3. Section 4 represents the proposed DT with block diagram. Section 5 describes the experimental analysis of proposed method with its results. The conclusion is described with future work in Section 6.

II. LITERATURE REVIEW

Researchers have suggested several techniques for the TSA. In this scenario, brief evaluations of some important contributions to the existing literatures are presented.

Jianqiang, Zhao, Gui Xiaolin, and Zhang Xuejun, [11] developed a word embeddings method based on large twitter data with the help of unsupervised learning by combining co-occurrence statistical characteristic and latent contextual semantic relationships between words in tweets. The sentiment features of tweets were formed by combining the word sentiment polarity score and n-gram features in word embeddings. The sentiment classification labels were predicted by feature set which was integrated into Deep Convolution Neural Network (DCNN). The efficiency of word embedding method was validated by conducting experiments on five datasets when compared with existing techniques. The pre-trained word vectors used in DCNN had good performance in the task of TSA. While clustering the sentimental contents in large dataset, the computational time becomes a bit high.

M.Z. Asghar, F. M. Kundi, S. Ahmad, A. Khan, and F. Khan, [12] proposed a hybrid classification framework to overcome the issues of incorrect classification. The performance of twitter-based SA systems were improved by using four classifiers such as a slang classifier, a moticon classifier, the SentiWordNet (SWN) classifier, and an improved domain specific classifier. The input text was passed through the first two classifiers such as emoticon and slag, after applying the preprocessing stage. In the final stage, SWN based and domain specific classifiers were applied to classify the text accurately. A limitation of the approach was the lack of automatic scoring of domain specific words without performing a lookup operation in SWN, which may increase the classification accuracy.

H. Ameur, et al., [13] proposed a Contextual Recursive Auto-Encoders "CoRAE" for predicting the polarity of sentiment label at sentence level and word level. The word vectors were reduced by using the Pointwise Mutual-Information-SA (PMI-SA) and CoRAE recursively combined each word with its neighbor's context word vectors by word order. The sentence vector representation was generated by using these continuous word representation. The efficiency of CoRAE was validated by using Sanders twitter corpus and Facebook comments corpus datasets. The method provides poor performance in classification accuracy when the initialization of parameters configuration of auto-encoders algorithm was less.

F. Bravo-Marquez, et al., [14] implemented a method for opinion lexicon expansion for automatically annotated tweets from three types of information sources such as tweets of emoticon-annotated, hand-annotated and unlabelled tweets. The domain-specific problem was tackled by transferring the method into annotation approach for unlabelled tweets. The disambiguated part-of-speech (POS) was included in the expanded lexicon for three polarity classes such as positive, negative and neutral. The linear relationship between sentiment and words were learned by using PMI-semantic orientation (PMI-SO) and stochastic gradient descent-SO (SGD-SO). Three datasets such as 6HumanCoded, Sanders and SemEval were used to validate the supervised lexicon frameworks. This approach used the labour-intensive approach to reduce the noise in labelled POS-disambiguated words, but it clean the data only by manually.

III. PROBLEM DEFINITION AND SOLUTION

This section describes a problem statement in sentiment analysis and also detailed about how the proposed system gives the solution to the problem statement.

Expert knowledge is required to select an appropriate classifier

After extracting the key-words from the pre-processed data, classification is carried out to classify the opinions of the customers for amazon products. In sentiment analysis, binary classifier like Support Vector Machine is a well-known classifier that is designed for the two-class problem. The success of binary classifier depends on the decision boundary that delivers good generalization performance. The major two problems accomplished in binary classifiers are ineffective in high dimensional data and only applicable for two-class classification. To address these issues, multiclass classification approaches are developed and analyzed their performance. Solution: In this research work, a new classifier: DT, MLP and Random Forest are implemented for classifying these tweets. The DT classifier effectively diminishes the size of resulting dual issue by developing a relaxed classification error bound. In addition, the undertaken classification approach quickly speeds up the training process by maintaining competitive classification accuracy.

IV. PROPOSED METHODOLOGY

Figure 1 shows Twitter sentiment classification framework using DT. By using this framework, it can obtain highly effective results for sentiment classification. The framework performs sentiment classification in four main phases: data collection using twitter datasets, twitter pre-processing incorporates filtering to filter unique twitter attributes, the polarity values of tweets are identify in pre-processing stage. Using Term frequency-inverse document frequency (TF-IDF) features, the features are extracted from the tweets for further classification process. The three categories of tweets are classified by initializing the Latent Dirichlet allocation (LDA), and then these categories are given as an input for decision tree for classifying tweets.

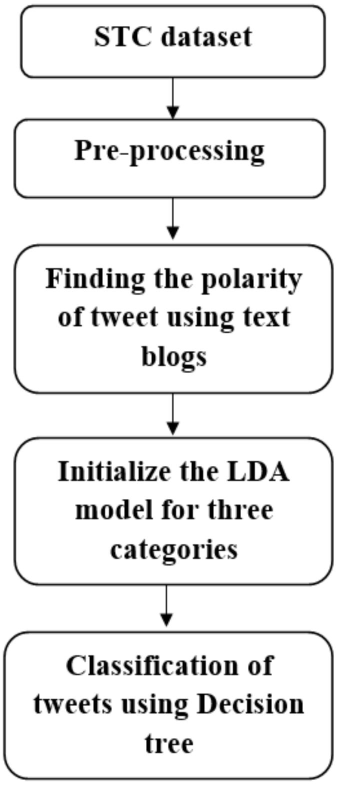
STC dataset
Pre-processing
Finding the polarity of tweet using text blogs
Initialize the LDA model for three categories
Classification of tweets using Decision tree

Fig 1. Block diagram of proposed decision tree

After the collection of data from STC dataset, data pre-processing is carried out in order to enhance the quality of collected data. The tweets had to go through the pre-processing steps prior to the classification because the language of Twitter has some unique attributes that may not be relevant to the classification process, such as usernames, links and hashtags. Figure 2 shows the sample image of twitter, which contains hashtags, informal dialects are removed.

FacilitiesManagement
@westernuFM

We've updated our twitter tag to @westernuFM to make finding us easier ...although, you have already found us. #bestfollowersever #westernu

9:44 AM - 25 Jul 2013

2 RETWEETS 1 FAVORITE

Fig 2. Sample image of twitter

In this proposed method, the pre-processing stage is carried out in two stages. In the first stage, the raw data contains more noise in terms of stop-words, and URLs, which are all removed effectively from the collected data. In second stage, lemmatization, the removal of stop words, lowercase conversion, and the stemming process were carried out to complete the twitter pre-processing.

Lemmatization: It transforms the words of a sentence into dictionary form. In order to extract the proper lemma, it is essential to analyse each morphological word. An example of lemmatization is denoted in table 1.

Table 1. A sample example of lemmatization

Form	**Morphological information**	**Lemma**
Studying	Gerund of the verb study	study
Studies	Singular number, third person, present tense of the verb study	study

The output of the pre-processing is the sequence of string which can be given as input for feature extraction. But, before extracting the useful information, the polarity of tweets can be identified. The existing methods find the polarity of tweets after the extraction of features which leads to poor performance. To improve the classification accuracy, the proposed DT used the textblog method to identify the polarity of texts.

Feature Extraction

In this section, a new feature extraction technique (term frequency-inverse document frequency) is undertaken for extracting the features from text data. The counter vectorization is a theory for representing or extracting the contextual words from a large corpus of text. Here, a set of mutual constraints is assign to each and every words for determining the similarity between the available words. Similarly, term frequency and inverse document frequency are accomplished to extract the useful features from text data. Term frequency and inverse document frequency measures how frequently a term appears in a document by using the equations (1) and (2).

$$Term\ frequency = \frac{no.of\ terms(t)apperars\ in\ document}{Total\ no.of\ yerms\ in\ a\ document} (1)$$

$$Inverse\ document\ frequency = log\frac{Total\ no.of\ documentt}{No.of\ document\ with\ term(t)} (2)$$

The technique TF-IDF is used for extracting the useful information for tweets. In next steps, the Latent Dirichlet allocation (LDA) method is initialized for categorizing the tweets into three types such as positive, negative and neutral. In this process, total 19 features are extracted from whole tweets which can be classified as three classes.

Latent Dirichlet Allocation

After extracting the information from the collected data, LDA approach is used for recognizing the tweets into three classes. Generally, the LDA is a probabilistic topic model, while each document is denoted as a random mixture of latent topics. Each and every latent topic is labelled as a distribution over a fixed set of words in LDA, which is utilized to identify the underlying latent topic structure on the basis of observed data. Usually, the words are generated in a two-phase process for every document. In the first phase, a distribution over topics is randomly selected for every document. In LDA, a word is a distinct data from a vocabulary index{1,..,V}, a series of N words are W.

The LDA is observed for the three-layered representation, π and μ parameters are examined during the generation of a corpus. For each and every document, the document-level topic variables are investigated. Respectively, the word level variables are examined for every word of the document in LDA. A joint distribution over random variable is represented as the generative process of LDA. The probability density function of the k-dimensional dirichlet random variable is determined by using the Eq. (3). Successively, the joint distribution of a topic mixture and the probability of a corpus are evaluated by using the Eq. (4) and (5).

$$p(\aleph|\pi) = \frac{\Gamma\left(\sum_{i=1}^{k}\pi_i\right)}{\prod_{i=1}^{k}\Gamma(\pi_i)}\aleph_1^{\pi_1-1}\ldots\ldots\aleph_k^{\pi_k-1}(3)$$

$$p(\aleph, x, y|\pi, \mu) = p(\aleph|\pi)\prod_{n=1}^{N}p(x_n|\aleph)\,p(y_n|x_n, \beta)\ \ (4)$$

$$p(D|\pi, \mu) = \prod_{d=1}^{M}\int p(\aleph_d|\pi)\times\prod_{n=1}^{N_d}\sum_{x_{dn}}(p(x_{dn}|\aleph_d)p(y_{dn}|x_{dn}, \mu)\,d\aleph_d(5)$$

Where ℵ is represented as the document-level topic variables, π is indicated as the dirichlet parameter, M is represented as the document, N is characterized as the number of words, μ is denoted as the topics, x is indicated as the per-word topic assignment, and y is indicated as the observed word.

In a document, the calculation of the posterior distribution of the hidden variable is an important inferential task of the LDA model. The exact inference of the posterior distribution of the hidden variable is an intractable problem. The combination of LDA with approximation algorithms like Gibbs sampling, Markov chain, Laplace, and variational approximation are extensively utilized for keyword extraction. The neutral, negative and positive classes are extracted with an individual weight value and these values are stored in dictionary. In order to attain neutral, negative and positive weight value, the testing sentiment data are coordinated with the dictionary in the testing phase. The values of three classes such as 30.86 for positive, 59.12 for neutral and 10.013 for negative tweets from collected data using LDA. After obtaining the neutral, negative and positive weight values, the classification process is carried out by using the decision tree algorithm.

Decision Tree Algorithm

Decision tree algorithm is a data mining induction technique that recursively partitions a data set of records using depth-first greedy approach or breadth-first approach until all the data items belong to a particular class. A decision tree structure is made of root, internal and leaf nodes. The tree structure is used in classifying unknown data records. At each internal node of the tree, a decision of best split is made using impurity measures. The tree leaves is made up of the class labels which the data items have been group. Decision tree classification technique is performed in two phases: tree building and tree pruning. Tree building is done in top-down manner. It is during this phase that the tree is recursively partitioned till all the data items belong to the same class label. It is very tasking and computationally intensive as the training data set is traversed repeatedly. Tree pruning is done is a bottom-up fashion. It is used to improve the prediction and classification accuracy of the algorithm by minimizing over-fitting (noise or much detail in the training data set). Over-fitting in decision tree algorithm results in misclassification error. Tree pruning is less tasking compared to the tree growth phase as the training data set is scanned only once. In the proposed system, the decision tree classification provides a better option for the end user to classify the positive and negative tweets. It is done by comparing the maximum frequent items generated by the rules in the training data have been compared with the maximum frequent items of the test data and hence the classification can be made easily.

This section detailed about the experimental result and discussion of the proposed system and also detailed about the performance metric, experimental setup, quantitative analysis and comparative analysis. The proposed system was implemented using Python with 4 GB RAM, 1 TB hard disk, 3.0 GHz Intel i5 processor. The performance of the proposed system was compared with other classification methods and existing research papers based on twitter dataset in order to assess the effectiveness of proposed system. The performance of proposed system was evaluated in terms of precision, recall, classification accuracy, and f-measure.

Performance measure

The performance measure is defined as the regular measurement of outcomes and results that develops reliable information about the effectiveness of the proposed system. Also, the performance measure is the process of reporting, collecting and analyzing information about the performance of a group or individual. Confusion metric to evaluate classifier for binary data is shown in Table 2 and could be understood in the following terms:

Table 2. Confusion Matrix.

	Predicted positive	**Predicted Negative**
Actual positive	True Positive (TP)	False Positive (FP)
Actual Negative	False Negative (FN)	True Negative (TN)

True positive (TP) represents the actual positive instances which are classified correctly as positive whereas false positive (FP) represents actual positive incorrectly classified as negative. Similarly, true negatives (TN) are actual negative instances and also correctly classified as negative and false negatives (FN) are actual negative instances and incorrectly classified as positive.

The mathematical equation of accuracy, f-measure, precision, and recall are denoted in the Eq. (6), (7), (8), and (9).

$$Accuracy = \frac{TN+TP}{TP+TN+FN+FP} \times 100 \qquad (6)$$

$$F - measure = \frac{2TP}{(2TP+FP+FN)} \times 100 \qquad (7)$$

$$Precision = \frac{TP}{(FP+TP)} \times 100 \qquad (8)$$

$$Recall = \frac{TP}{(FN+TP)} \times 100 \qquad (9)$$

Where, TP is signified as true positive, TNis indicated as true negative, FP is specified as false positive, and FN is indicated as false negative.

Performance Analysis of Proposed Decision Tree

In this experimental research, simulated twitter data is used for comparing the performance evaluation of existing methodologies and the proposed approach in terms of accuracy, F-measure, precision and recall. Two series of experiments under various experimental circumstances are conducted in STC datasets. The experiments primarily aim to determine the best supervised approach when compared with random forest (RF), MLP and DT. The DT is calculated for each of the topic-based text classified tweet of Sanders corpus. This classification helps the organization in focusing on tweets with highest impacts. The discussion of DT of Sanders defined topics in subsequent sections are represents below.

Performance Analysis in terms of Precision, Recall and F-Measure

In this section, the performance of DT is compared with MLP, RF in terms of precision, recall and F-measure for three classes such as positive, negative and neutral in Table 3. The overall performance of precision, recall and F-measure of DT algorithm with existing techniques such as RF and MLP in Table 4.

Table 3. Performance of DT for three classes of STC dataset

Methods	**Parameters**	**Positive**	**Negative**	**Neutral**
	Precision	0.56	0.27	0.66

RF	Recall	0.39	0.04	0.86
	F-Measure	0.46	0.07	0.74
MLP	Precision	1.00	1.00	0.59
	Recall	0.06	0.02	1.00
	F-Measure	0.11	0.04	0.75
Proposed DT	Precision	**0.92**	0.46	0.86
	Recall	0.74	0.46	**0.95**
	F-Measure	0.82	**0.46**	0.91

In the above table, the best values of proposed DT for three classes is marked as bold. The DT achieved 92% precision for positive class, 46% for negative and 86% for neutral, whereas the best recall value for DT is 95% for neutral, 46% for negative and 74% for positive. The F-Measure of DT is 82% for positive, 91% neutral and 46% for negative. While compared with existing techniques, MLP provides poor performance on negative class of recall. The RF method achieved highest neutral value is 66% of precision, 86% of recall and 74% of F-Measure.

Table 4. Overall performance of DT in terms of precision, recall and F-Measure

Methods	Precision	Recall	F-Measure
RF	81.16	80.85	80.83
MLP	74.74	74.20	74
Proposed DT	**88.12**	**92.31**	**90.17**

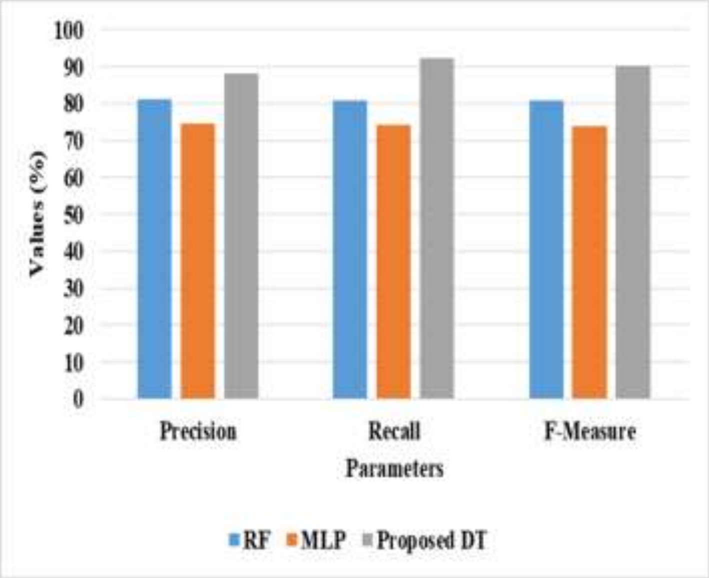

Fig 3. Performance of DT

From the above table 4, it is clearly showing that the proposed DT achieved higher values of precision, recall and F-Measure when compared with existing techniques. But the RF achieved 81.16% precision, 80.85% recall and 80.83% F-Measure by using weighting approach. The precision value of MLP is 74.74%, whereas the recall value is 74.20%. The next section will describe the performance of DT in terms of accuracy.

Performance Analysis of DT in terms of accuracy

In this section, the classification accuracy of proposed DT are compared with existing techniques such as RF and MLP. Table 5 represents the accuracy values of proposed method with existing system.

Table 5. Performance of Accuracy for proposed DT

Methods	**Accuracy**
RF	56.46
MLP	80.97
Proposed DT	**99.51**

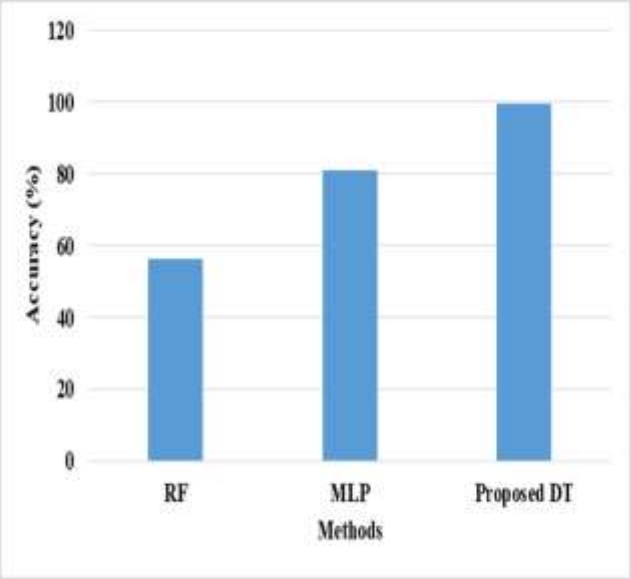

Fig 4. Performance of DT in terms of Accuracy

From the experimental results, the proposed DT achieved higher accuracy (i.e 99.51%) for twitter classification from STC datasets. The RF achieved very low accuracy when compared with all other existing techniques. The MLP approach obtained 80.97% accuracy, but failed to focus on computation work amount. The testing results of the STC dataset is compared with the training results. The best performances of the DT method are produced in STC dataset when compared with existing techniques such as RF and MLP. It could be seen that the number of samples in the testing and training datasets played a significant role in producing highly accurate results.

The results indicate that the proposed DT proves its capability to enhance the discriminating power of terms for tweet classification. The proposed method can aid in dimensionality reduction, sentiment analysis, spam detection, and many other applications that face different challenges of tweet. Therefore, the proposed scheme mitigates the effect of these challenges on the performance of the classification task. Although our results have been achieved for data from Twitter, one of the most common social media platforms, we believe that this proposed method is also applicable to other social media.

V. CONCLUSION

TSA is one of the emerging research fields for analyzing and identifying the sentiments and viewpoints of users. The proposed method consists of two stages such as pre-processing and classification of tweets. The acquired twitter data is pre-processed by eliminating the unnecessary emoji from the tweets, and the execution of missing value treatment. The pre-processed data is utilized for TSA using DT algorithm. The experimental investigation of

DT is verified on simulated Sanders twitter data, which showed the superiority of the proposed approach. The classification rate of Sander twitter data is better in the proposed methodology than the previous methodologies. Compared to other existing approaches in TSA, the proposed scheme delivered an effective performance by means of accuracy, F-Measure, precision and recall. The developed approach improved the classification precision and recall rate of around 3-15% compared to the previous methods. In future work, to improve the classification rate, a hybrid sentiment approach will be developed for other social media data such as youtube and facebook to identify the sentiment of people towards certain issues.

VI. REFERENCES

[1] Luis Terán, and José Mancera, "Dynamic profiles using sentiment analysis and twitter data for voting advice applications."Government Information Quarterly (2019).

[2] H. Shirdastian, M. Laroche, and M.O. Richard, “Using big data analytics to study brand authenticity sentiments: The case of Starbucks on Twitter,” International Journal of Information Management, 2017.

[3] Himja Khurana, and Sanjib Kumar Sahu. "Bat inspired sentiment analysis of Twitter data."Progress in Advanced Computing and Intelligent Engineering. Springer, Singapore, pp. 639-650, 2018.

[4] M. Daniel, R.F. Neves, and N. Horta, “Company event popularity for financial markets using Twitter and sentiment analysis,” Expert Systems with Applications, vol.71, pp.111-124, 2017.

[5] Y. Ruan, A. Durresi, and L. Alfantoukh, “Using Twitter trust network for stock market analysis,” Knowledge-Based Systems, vol.145, pp.207-218, 2018.

[6] R.C. LaBrie, G.H. Steinke, X. Li, and J.A. Cazier, “Big data analytics sentiment: US-China reaction to data collection by business and government,” Technological Forecasting and Social Change, 2017.

[7] M. Komorowski, T. Do Huu, and N. Deligiannis, “Twitter data analysis for studying communities of practice in the media industry,” Telematics and Informatics, vol.35, no.1, pp.195-212, 2018.

[8] V. Vyas, and V. Uma, “An Extensive study of Sentiment Analysis tools and Binary Classification of tweets using Rapid Miner,” Procedia Computer Science, vol.125, pp.329-335, 2018.

[9] T. Singh, and M. Kumari, “Role of text pre-processing in twitter sentiment analysis,” Procedia Computer Science, vol.89, pp.549-554, 2016.

[10] M. Birjali, A. Beni-Hssane, and M. Erritali, “Machine Learning and Semantic Sentiment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks,” Procedia Computer Science, vol.113, pp.65-72, 2017.

[11] Jianqiang, Zhao, Gui Xiaolin, and Zhang Xuejun. "Deep convolution neural networks for Twitter sentiment analysis."IEEE Access vol. 6, pp. 23253-23260, 2018.

[12] M.Z. Asghar, F. M. Kundi, S. Ahmad, A. Khan, and F. Khan, “T-SAF: Twitter sentiment analysis framework using a hybrid classification scheme”, Expert Systems, vol. 35, no. 1, e12233, 2018.

[13] H. Ameur, Salma Jamoussi, and Abdelmajid Ben Hamadou. "A New Method for Sentiment Analysis Using Contextual Auto-Encoders."Journal of Computer Science and Technology 33.6 (2018): 1307-1319.

[14] F. Bravo-Marquez, Eibe Frank, and Bernhard Pfahringer. "Building a Twitter opinion lexicon from automatically-annotated tweets."Knowledge-Based Systems 108 (2016): 65-78.

[15] I. Alsmadi, and Gan Keng Hoon. "Term weighting scheme for short-text classification: Twitter corpuses."Neural Computing and Applications (2018): 1-13.

[16] H. S. Manaman, S. Jamali, and A. AleAhmad. "Online reputation measurement of companies based on user-generated content in online social networks."Computers in Human Behavior 54 (2016): 94-100.