INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 3 - V7I3-2142)
Available online at: https://www.ijariit.com

Improved tweet Sentiment Classification Using Convolution Neural Network and Random Forest

Pallavi Sharma
professional.erpallavi@gmail.com
D.A.V. Institute of Engineering and Technology, Jalandhar, Punjab

Dr. Harpreet K. Bajaj
harpreet_daviet@yahoo.in
D.A.V. Institute of Engineering and Technology, Jalandhar, Punjab

ABSTRACT

With over 319 million monthly active users, Twitter has developed into a goldmine for organizations and people with a strong political, social, or economic incentive to retain or enhance their clout and reputation. Sentiment analysis enables these firms to conduct real-time surveys on numerous social media platforms. Twitter sentiment analysis technology enables the measurement of public attitude toward certain events or products. The majority of current research is devoted to extracting sentiment traits through the analysis of lexical and syntactic variables. These characteristics are openly stated using emotional words, emoticons, and exclamation points, among others. In this research, effective feature extraction is accomplished via the use of convolution mapping and an attention layer. These features are then learned by random forest.

Keywords— *Convolution Neural Network, Twitter*

1. INTRODUCTION

Sentiment analysis is a process which deals to identifying the feelings, emotions, attitude of a person towards some product, entity, or any political and personal issue. Sentiment analysis is a method to extract the hidden information from the web data. This hidden information gives a lot of information related to user's view and opinion. Sentiment analysis performed on the basis of subjective and objective nature of text. The subjective data defines the sentiment part of the text and rest part that is objective part does not contain sentiment information in it. The most important part of the sentiment analysis is extracting the features from the text and classify the data according to these features. Here is an example of subjective and objective data.
Subjective: The flavor of appy fizz is best and it is like original apple.

1.1 Objective: I taste this last Sunday

On the basis of this text is also divided into three categories positive, negative, neutral which express its view or opinion. The rapid generation of web it leads to a huge amount of data generated by the user. There are different kinds of users which interact to each other on web and generate the content in various forms like:

- **Weblogs:** Weblogs are basically a collection of blogs and their list mainly in the form of hyperlinks which are used to recommend the other website of blog.
- **News:** These are the daily events and activities happened around the world. The news also contains the new research and discovery all over the world.
- **Reviews:** These are the feedback from the users of product or services from the web. The reviews are mostly used in the e-commerce and entertainment industry to give the feedback and feeling related to the services of the e-commerce company.
- **Social Networking Sites:** the biggest boom in the technology is social media platforms like Facebook, Twitter, and Google +, etc. on these platforms users text each other's, share their day to day events and also mention about their special days.

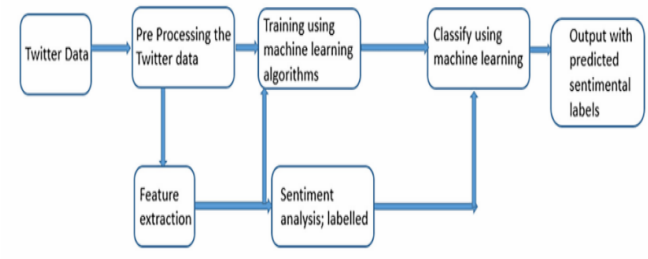

Figure 1: Basic method of sentiment analysis

In the era of internet and technology the e-commerce growing with very high rate because it provides an effective platform to the customers and manufacturers to interact with each other. The reviews of the product plays and important role in its overall sale and it depend on its quality. The review helps to the new consumers in deciding whether the product is good or not. It also helps the manufacturer to know about the status of the product in the market after its sale. This thing helps to improve the product sale by quantitatively as well as qualitatively. Tweet analysis according to its sentiment is a challenging process

because of the huge number of tweets and the variation in tweets. The aim of this research is to effectively analyze with volume and variation of tweets. The process of analyzing variation in tweets reduces their accuracy, so layer by layer refinement or monitoring is an essential parameter for effective analysis of tweet. The proposed methodology follows the aims and objectives of the research by using tweet text preprocessing by TF-IDF (Term Frequency- inverse document frequency) for obtaining domain knowledge of text and for classifier model us deep learning (deep neural networks). By this effective approachable to analysis tweet sentiment by approximate human generated accuracy. Analyzing the public sentiment is important for many applications such as firms trying to find out the response of their products in the market, predicting political elections and predicting socioeconomic phenomena like stock exchange.

1.2 Different Types of Sentiment Analysis

Sentiment analysis of the data can be done by using different ways at different place. The analysis based on the document, sentence, and feature level. The below given section give the brief details on it.

1.2.1 Sentence Level analysis: This analysis based on the sentiment determines by one sentence and determine the sentiment is positive, negative, and neutral. This analysis is mainly related to the subjective classification and defines the subjective views and sentiments from it. Eg. This place is magnificent and impressive. This sentence defined the positive review against some place and it is decided by the words in it that are impressive and magnificent

1.2.2 Document Level Analysis: Document level sentiment analysis determines the sentiment of the complete document is in positive or negative. The review of a product defined the complete document sentiment and it is applicable on the single entity level document. This type of analysis does not applicable on the multiple entities.

1.2.3 Feature Level Analysis: The document level analysis and sentence level analysis determine the sentiments at different levels but not able to defines the exact sentiment. This issue solved by using the feature level sentiment analysis on the basis of different features of the single entity. For example, review of a restaurant can be determined on the different features like their service, taste, food quantity, and charges.

1.3 Application of Sentiment Analysis

Sentiment analysis is mainly used to understand the subjective nature of a text. Following are the different areas where sentiment analysis is used.

- **Decision Making:** Decision making is an important part of our life in which we take some decision on different topics related to day-to-day life. The sentiment analysis helps in e-commerce field to decide which product is better and which is not on the basis of sentiments from others users.
- **Designing and building innovative products:** Sentiment analysis plays a vital role in the industrial growth because it helps to find the product review from the consumers. The reviews collected from the consumers define the quality and quantity of the product and helps to improve the product quality if needed.
- **Recommendation System:** these systems are based on the previous data of the user which contains its previous likes and dislike products, and personal information. The system uses the data mentioned before and gives some suggestion for the product selection.
- **Product Analysis:** Sentiment analysis helps to analyze the product on the basis of reviews. It helps in product selection on the basis of its feature.
- **Business Strategies:** Business strategies are made on the basis of response from the peoples and their reviews. This is required because all the business in the world connected through the technology.

2. RELATED WORK

Wagh et al. (2020) By combining the natural language toolkit with machine learning, we suggested an approach for sentiment analysis. NLTK is mostly used for data approach and classification programming. The suggested methodology was utilised to estimate the polarity of sentiment in user-generated data. Twitter data was acquired in this study using the Twitter API, and then pre-processed using stop word removal, stemming, and lemmatization. The next step of the task involves classifying the data features using the Naive Bayes and multinomial classifiers. Precision, recall, accuracy, and f-measure were used to evaluate performance.

Asghar et al (2020) presented the hybrid classification approach for the sentiment analysis of the twitter data. In this study, four different classifiers are used for the classification and solve the issues related to the classification of data. The classifiers used in this study were slang, emoticon, senti word net, and domain specific classifier. The working of this model starts from the preprocessing of input data and then applies the slang and emoticon classifier. After this rest two classifiers are applied to effective classification. The classification based in this work is on document and sentence level. The result of this study reveals that is resolves the issue by considering slang and emoticons.

Fouad et al (2020) Presented a sentiment analysis methodology based on the processing of natural languages and data. The machine learning idea is used to recognize good and negative tweets. Various approaches in the training phase for the input tweets are employed with different characteristics. The model delivers users' views with its basic tweets.

Asghar, Muhammad et al. (2019) built a supervised white-box sentiment analysis system for microblogging using the rule induction methodology. The studies are evaluated via the lens of sough set theory; a field of mathematics founded on induction procedures. Classification rules are developed by training using decision tables and induction algorithms. RST classifies tweets as either positive, negative, or neutral. Learning is performed in this manner via the use of LEM and tweet-based corpus-based rules. The accuracy, coverage, and number of rules employed are all utilized to quantify the performance.

Garcia et al. (2019) Presented a novel strategy for improving the accuracy of decision support systems in the sentiment analysis approach of probabilistic classifiers. Information processing is automated via the use of the DSocial Platform and enhances the accuracy of sentiment analysis support systems. In this model, the major task is to implement the probabilistic classification. The approach helps enhance user prediction based on their sense of a movie.

Chidananda et al. (2019) proposed the N-gram method for sentiment analysis. The proposed model based on the combination of N-gram and log function to find sentiment from the twitter data and provides the robust results with better accuracy. This model based on the data of twitter and identifies the opinion based on the tweets. In new model same word

considered as positive review but in previous work it considered as negative review.

Alrubaian et al. (2019) presented a model for the credibility analysis of the information from the twitter analysis. The model basically consists of the four basic components that are reputation-based credibility classifier, user experience, and a feature ranking algorithm. All the components work together to analyze and asses the credibility of tweeter data. The performance evaluation of the model tested on the two different test and 10-fold cross validation. The precision and recall of the model result is effective and efficient.

Pujari et al. (2018) introduced a new framework which consist of built-in packages of python and mine the customer's views related to the product and make groups according to the sentiments. The classification review done on the basis of three classification algorithms that were SVM (support vector machine), naïve bayes (NB), and maximum entropy. This model provides the effective results in sentiment analysis and easily to extend with other technology for more effective results.

Sood, et al. (2018) Enabled an exploration of the effects of social media and news on the individual, their perspective, and their own identity. This effort is done to understand the thoughts and feelings behind depression with R studio, which extracts tweets. Depression state is detected by the algorithm which recognises the Sentiments, and the score is then awarded to each sentiment using which, depression state. Additionally, this project assists in identifying persons with mental illness, which harms society and the individual.

Yu, et al. (2018) the academic track record of students has been used to build a model that can predict who would fail in the coursework. The work is completed on the basis of evaluations from the pupils, who gave them in the form of written remarks. The text's sentiment is classified using the convolution neural network classifier and SVM (support vector machine) classifier. In the subsequent prediction classification, the confusion matrix is created after this. This model's performance is gauged by three metrics: precision, recall, and accuracy. The T-test is performed on the data to do the statistical analysis. Students will have increased control and independence in the learning process using this strategy.

Zheng et. al. (2018) compared the online health support groups' function in assisting patients and their carers to that of face-to-face support groups. This project's goal is to raise awareness and provide support for those with sickle cell anaemia.

To gather data for this purpose, the Facebook graph API is used to collect their comments and Facebook groups' information. The supervised machine learning algorithm (SVM) is used after this to assess the data set and generate labels for the comments, each of which categorises the data set into three groups according to the positive, negative, and neutral states. The categorization of the messages indicates that this methodology successfully catalogues the communications.

3. PROPOSED METHODOLOGY

A social network and sentiment analysis framework has been developed that can be run on Twitter data to study work. Twitter is a popular social networking and microblogging site that counts hundreds of millions of active users and regular posts. As a social networking site, Twitter is designed as a guided graph that allows individual users to monitor a number of other users (followers) and other users (followers) to do the same. In proposed approach tweet classified using classifier which efficiently classified tweets according to its sentiment.In proposed approach hybridization Convolution based features with Random Forest.

3.1 Convolution Neural Networks

Convolutional Neural Networks (CNN), were first introduced by Yann LeCun's in 1998 for Optical Character Recognition (OCR), where they have shown impressive performance on character recognition. CNN is not just used for image related tasks, they are also commonly used for signals and language recognition, audio spectrograms, video, and volumetric images.

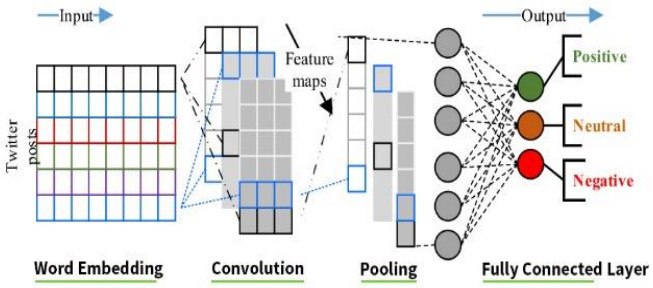

Figure 2: CNN uses multiple layers in its architecture

Following are the layers used to build convolutional neural network architectures.

- *Convolutional Layer*
- *Activation Layer*
- *Pooling Layer*
- *Fully-Connected Layer or Densely Connected Layer*
- *Output Layer or SoftMax Layer for classification*

Convolutional Neural Networks (CNN), were first introduced by Yann LeCun's in 1998 for Optical Character Recognition (OCR), where they have shown impressive performance on character recognition. CNN is not just used for image related tasks, they are also commonly used for signals and language recognition, audio spectrograms, video, and volumetric images.

3.2 Random Forest

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction (see figure below).The low correlation between models is the key. Just like how investments with low correlations (like stocks and bonds) come together to form a portfolio that is greater than the sum of its parts, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions.

The reason for this wonderful effect is that the trees protect each other from their individual errors (as long as they don't constantly all err in the same direction). While some trees may be wrong, many other trees will be right, so as a group the trees are able to move in the correct direction.

So the prerequisites for random forest to perform well are:

1. There needs to be some actual signal in our features so that models built using those features do better than random guessing.
2. The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

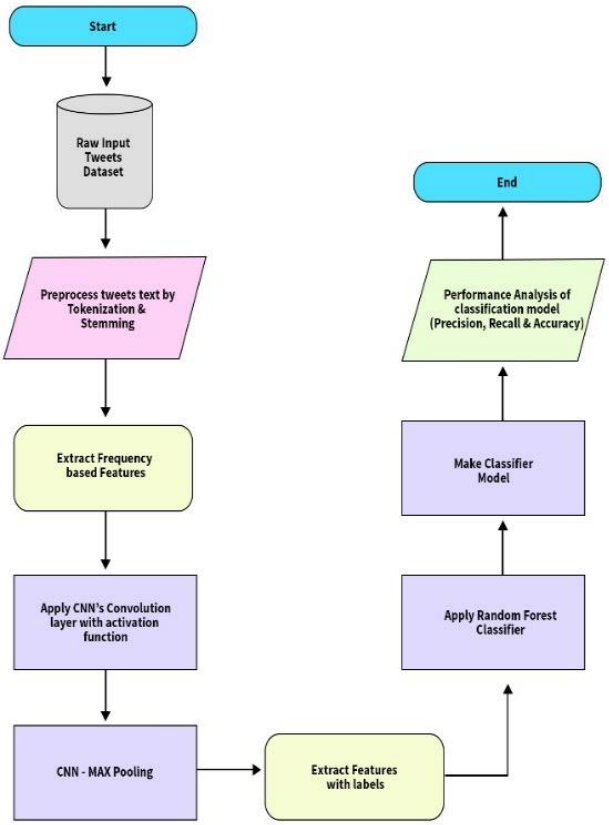

Figure 3: Flowchart for active anchor node selection.

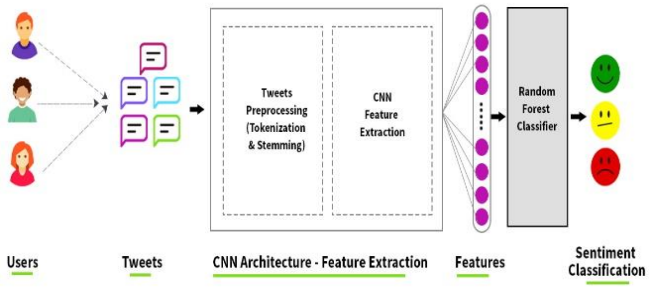

Figure 4: Proposed CNN-RF architecture of hybrid classification model

It is gathered from data set for experimentation and are stored in the preprocessing database(**https://www.kaggle.com/c/tweet-sentiment-extraction/data**). In the section below, the proposed model methodology and the techniques used in this work are listed in detail.

Stage 1: Collection of Data

From the tweet's analysis, the data provided as an input to the proposed model were collected. In order to extract useful information from twitter data set, several preprocessing steps are needed.

Stage 2: Data storage and compilation

The tweets collected are saved to format .csv files, which are then fetched to Python. There are about 3000 tweets available to train the datasets and test them.

Stage 3: Pre-processing of data

The twitter data can be cleaned in this phase. Preprocessing Tweets Analysis extracts the noisy and repetitive data from raw data and then generates the qualified data collection to work further.

Various steps are taken to clean the following details.

i. The entire uppercase is transformed into a reduced one.

ii . ii. Delete from the data all internet slangs.

iii. iii. Delete from the list all stopwords.

iv. Delete all other white spaces.

v. Duplicate words are compressed.

vi. All hash tags are deleted, but the text of the hash tag is reserved.

Stage 4: Frequency base Features and Convolution Neural network

In This step extract frequency base features. In frequency base features extract frequency of unique word which present in tweets after features extraction apply convolution layers with pooling layer and get transform feature vectors.

Stage 5: Apply Classifier

In this step apply random forest on transform feature vector and label and make classifier model and analysis performance metrics

3.3 Performance Parameters

- Accuracy Classification based on accuracy is generally what we mean when we use the word accuracy. This represents the measure of the number of appropriate predictions to total corresponding inputs. The proposed and the existing model are comparable on the basis of accuracy. It only functions well when the number of observations belonging to every class is equal.
- **Precision:** described as TP/ (FN+TP). equivalent to the number of positive data sources that are accurately considered to be positive for all the positive data-based points.
- **Recall** described as TN / (FP+TN). recall relates to the ratio of negative database points correctly deemed negative in relation to all negative data points.
- **F**-score **F-score** or **F-measure** is a measure of a test's accuracy. It is calculated from the precision and recall of the test, where the precision is the number of correctly identified.

4. EXPERIMENT AND RESULTS

Table 1: comparison of Proposed and existing approaches in training

Approach	Accuracy (training)	Precision (training)	Recall (training)	F-score (training)
CNN	90	96.2	92	93
CNN-RF	91.2	94.12	92.13	94

In table 1 and figure show the comparison between training time different performance metrics in proposed (CNN-RF) and existing CNN approach both approaches perform well but average performance improvement in proposed approach.

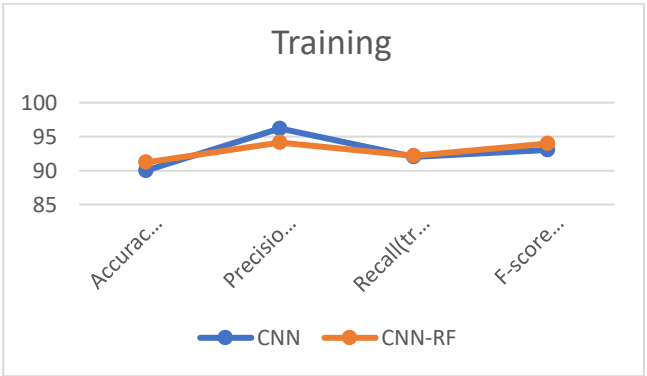

Figure 5: comparison of Proposed and existing approaches in training

Table 2: comparison of Proposed and existing approaches in Existing

Approach	Accuracy (Testing)	Precision (Testing)	Recall (Testing)	F-score (Testing)
CNN	83.2	72	69.12	72.12
CNN-RF	87	84	75.12	76

In table 1 and figure show the comparison between testing time different performance metrics in proposed (CNN-RF) and existing CNN approach both approaches perform well but average performance improvement in proposed approach.

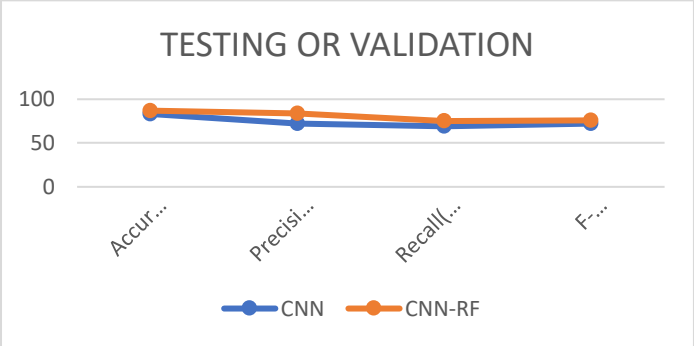

Figure 6: Comparison of Proposed and existing approaches in existing

5. CONCLUSION

To improve the accuracy and speed of analysis, we trained a deep neural network using a convolution approach using Twitter sentiment analysis. To begin, global vectors for word representation are learned, which act as features for the word sentiment information. Following that, we concatenate these word representations with the prior polarity score feature and state-of-the-art features to generate an attention layer feature set. These feature sets are combined and fed into a deep convolutional neural network, which is used to train and predict sentiment classification labels for tweets. A deep convolutional neural network is capable of successfully constructing text semantics. Its objective is to address the issue of data scarcity. Convolutional neural networks with random forest successfully predict context sentiment directly from text and detect the most critical components of a tweet. It eliminates error propagation and increases classification performance greatly.

6. REFERENCES

[1] Wagh, Bhagyashri, J. V. Shinde, and P. A. Kale. "A Twitter Sentiment Analysis Using NLTK and Machine Learning Techniques." *International Journal of Emerging Research in Management and Technology* 6.12 (2020): 37-44.

[2] Asghar, Muhammad Zubair, et al. "T-SAF: Twitter sentiment analysis framework using a hybrid classification scheme." *Expert Systems* 35.1 (2020): e12233.

[3] Fouad, Mohammed M., Tarek F. Gharib, and Abdulfattah S. Mashat. "Efficient Twitter Sentiment Analysis System with Feature Selection and classifier Ensemble." *International Conference on Advanced Machine Learning Technologies and Applications*. Springer, Cham, 2020.

[4] Asghar, Muhammad Z20142ubair, et al. "RIFT: A Rule Induction Framework for Twitter Sentiment Analysis." *Arabian Journal for Science and Engineering* 43.2 (2019): 857-877.

[5] García-Díaz, Vicente, et al. "An approach to improve the accuracy of probabilistic classifiers for decision support systems in sentiment analysis." *Applied Soft Computing* 67 (2019): 822-833.

[6] Chidananda, Himadri Tanaya, Debashis Das, and Santwana Sagnika. "Sentiment Analysis Using N-gram Technique." *Progress in Computing, Analytics and Networking*. Springer, Singapore, 2019. 359-367.

[7] Alrubaian, Majed, et al. "A credibility analysis system for assessing information on twitter." *IEEE Transactions on Dependable and Secure Computing* 15.4 (2019): 661-674.

[8] Pujari, Chetana, and Nisha P. Shetty. "Comparison of Classification Techniques for Feature Oriented Sentiment Analysis of Product Review Data." *Data Engineering and Intelligent Computing*. Springer, Singapore, 2018. 149-158.

[9] Sood, Akriti, "An Initiative to Identify Depression using Sentiment Analysis: A Machine Learning Approach." Indian Journal of Science and Technology,2018, 11.4.

[10] Yu, L. C., "Improving early prediction of academic failure using sentiment analysis on self-evaluated comments." Journal of Computer Assisted Learning, 2018.

[11] Zheng, K., Li, A., & Farzan, R., "Exploration of Online Health Support Groups through the Lens of Sentiment Analysis." International Conference on Information, Springer, 2018, pp. 145-151.

[12] Batra, R., & Daudpota, S. M. "Integrating StockTwits with sentiment analysis for better prediction of stock price movement." In Computing, Mathematics and Engineering Technologies (iCoMET), IEEE, 2018, (pp. 1-5).

[13] Rahman, L., Sarowar, G., & Kamal, S. (2018). Teenagers Sentiment Analysis from Social Network Data. In Social Networks Science: Design, Implementation, Security, and Challenges (pp. 3-23). Springer, Cham.

[14] Mehra, R., Bedi, M. K., Singh, G., Arora, R., Bala, T., & Saxena, S. (2017, July). Sentimental analysis using fuzzy and naive bayes. In Computing Methodologies and Communication (ICCMC), 2017 International Conference on (pp. 945-950). IEEE.

[15] Krishna, B. Vamshi, Ajeet Kumar Pandey, and AP Siva Kumar. "Feature Based Opinion Mining and Sentiment Analysis Using Fuzzy Logic." Cognitive Science and Artificial Intelligence. Springer, Singapore, 2018. 79-89.

[16] Shidaganti, Ganeshayya, Rameshwari Gopal Hulkund, and S. Prakash. "Analysis and Exploitation of Twitter Data Using Machine Learning Techniques." International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications. Springer, Singapore, 2018. 135-146.

[17] Rout, Jitendra Kumar, et al. "A model for sentiment and emotion analysis of unstructured social media text." Electronic Commerce Research 18.1 (2018): 181-199.

[18] Mumtaz, Deebha, and Bindiya Ahuja. "A Lexical and Machine Learning-Based Hybrid System for Sentiment Analysis." Innovations in Computational Intelligence. Springer, Singapore, 2018. 165-175.

[19] Al-Smadi, Mohammad, et al. "Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews." International Journal of Machine Learning and Cybernetics (2018): 1-13.

