Research Article

Genetic Algorithm Based Hybrid Model Ofconvolutional Neural Network And Random Forest Classifier For Sentiment Classification

Siji George C G[*1], B. Sumathi[2]
[1] Research Scholar, Department of Computer Science, CMS College of Science and Commerce, India
[2] Associate Professor, Department of Computer Science, CMS College of Science and Commerce, India

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: Sentiment analysis is one of the active research areas in the field of datamining. Machine learning algorithms are capable to implement sentiment analysis. Due to the capacity of self-learning and massive data handling, most of the researchers are using deep learning neural networks for solving sentiment classification tasks. So, in this paper, a new model is designed under a hybrid framework of machine learning and deep learning which couples Convolutional Neural Network and Random Forest classifier for fine-grained sentiment analysis. The Continuous Bag-of-Word (CBOW) model is used to vectorize the text input. The most important features are extracted by the Convolutional Neural Network (CNN). The extracted features are used by the Random Forest(RF) classifier for sentiment classification. The performance of the proposed hybrid CNNRF model is comparedwith the base model such as Convolutional Neural Network (CNN) and Random Forest (RF) classifier. The experimental result shows that the proposed model far beat the existing base models in terms of classification accuracy and effectively integrated genetically-modified CNN with Random Forest classifier.

Keywords: Convolutional Neural Network(CNN), Genetic Algorithm(GA), Random Forest(RF), Sentiment Analysis, Classification.

1. Introduction

Sentiment analysis is the natural language processing task which is used to identify the sentiment expressed in a particular document. This document may be a user comment/opinion regarding a particular product or service. Sentiment refers to the feeling which is come from within a review or comment. The process of finding the author attitude towards a particular piece of content with respect to a particular topic. This author attitude may be positive, negative, neutral or have no sentiment at all. The sentiment analysis focuses on emotions, feelings, urgency, polarity and even intensions. Depending on the objective to be met, the sentiment analysis can be defining to meet particular need. There are different types of sentiment analysis such as emotion detection, aspect-based sentiment analysis, fine-grained sentiment analysis, multi-lingual sentiment analysis etc.

Sentiment analysis is highly demanded because it helps to quickly identify the feeling or opinion about a particular service or product. Huge volume of unstructured data is generating daily through social conversations, chats, email, documents, surveys etc. These unstructured data can be used for different applications. But it is really tough to use for sentiment in an efficient and timely manner. Sentiment analysis also known as opinion mining. It is mainly works with machine learning algorithms and natural language processing (NLP) to automatically identify the emotional tone with in the online conversations. Based on the volume of data, it is possible to use different algorithms for performing sentiment analysis.

Machine learning algorithms are widely used for performing sentiment analysis. Depending on the objective, sentiment analysis algorithm can be performed at document-level, sentence-level or sub-sentence level. Document level analysis is done to find out the sentiment for the entire text. In sentence level sentiment analysis, it is try to identify the sentiment expressed with in each sentence in the document. Sub-sentence level sentiment analysis more complex and it perform the analysis of sentiment of sub-expression with in a sentence.

Both supervised and unsupervised machine learning algorithms are available for sentiment analysis. The supervised algorithms work based on set of labelled dataset. These algorithms will collect the data and will produce the output from the previous experience. It is widely used for solving real-world computational problems. In text classification, it is always not possible to create large volume of labelled data but it is simple to collect it. Unlike supervised algorithms, no teacher or supervisor is needed for the unsupervised machine learning algorithms. These algorithms are capable to perform analysis on unlabelled data. That is, the machine is fully restricted to identify the hidden structure with in the unlabelled data by itself.

In order to utilize the full potential of sentiment analysis tools, it can have plugged with deep learning models. Deep learning is a part of machine learning which uses the power of artificial intelligence to process the data like the human brain does. There are lot of deep learning algorithms which allow us to accurately handle huge volume of data with very little human support. Sometimes in machine learning, there are chances for making mistakes by machines and human must provide some input to correct it. However, in deep learning the neural networks are capable to learn and correct itself through its algorithm chains. The initial training stage of

deep learning is time consuming until it starts to learn on its own. Once the neural networks are trained, they can solve more complex problems.

Most of the researchers are use neural networks and machine learning models separately. In this paper a hybrid model of Convolutional Neural Network with Random Forest supervised machine learning classifier is used with genetic algorithm. The genetic algorithm helped to tune the hyperparameters of the proposed model.

The remaining sections of the paper is organized as follows. Section 2 explains the related work in the same field. Proposed methodology and architecture are given in Section 3. Section 4 gives the details of the dataset used and results of the experiments. Conclusion of the work is given in Section 5.

2. Related work

This section describes the applications of Convolutional Neural Network (CNN) in natural language field. Most recent works which used GA, grid search and random search for CNN parameter optimization are also discussed here.

An automatic classification method used to classify online judgement is developed by Neha Bansal et al. in 2019 [1].The authors used a hybrid approach of Conventional Neural Network(CNN) and bidirectional long short-term memory (BiLSTM). In this work, GA was used to get optimal word vector and this is given as input to BiLSTM with Softmax classifier. The experimental results show that the proposed model exceeds performance of existing individual models in terms of accuracy.

A new model is proposed by AshrayBhandre and DevinderKaur in 2018 [2] for classifying handwritten numbers. The authors used MNIST dataset for their experiments. GA was used to extract correct architecture for CNN by evolving correct hyperparameters for the given application.The proposed CNN model with GA achieved an accuracy of 99.2% in best out of 10 runs.

A variable length genetic algorithm was used by Xueli Xiao et al. in 2020 [3] for systematically tune the hyperparameters of CNN to improve the performance. A detailed comparison of different optimization methods such as random search, large scale evolution and classical genetic algorithms have done in this work. According to the authors, the more time spent on hyperparameter optimization will definitely result in higher performance.

A detailed comparison among optimization techniques such as random search, grid search and genetic algorithm is done by PetroLiashchynskyi et al. in 2019 [4]. The authors used these algorithms for designing the Conventional Neural Network. CIFAR-10 dataset with augmentation and pre-processing methods are used as the dataset for their work. According to the authors' experience the grid search is not suitable for large search space. The random search does not guarantee good result still it is little faster algorithm. The authors suggest the genetic algorithm when it have large search space and too many parameters to be optimized.

New version of random search for hyperparameter optimization is proposed by Adrian Catalin et al. in 2020 for machine learning algorithms [5]. This improved random search version generates new value for every hyperparameter with probability of change. The proposed random search version outperforms the standard random search method. The authors used this for optimizing the CNN hyperparameters. This can be used for any optimization problem in discrete domain.

3. Proposed methodology

The proposed system architecture is used to develop a classification model which is configured automatically by using genetic algorithms. This proposed system includes mainly three stages: word embedding, feature extraction using CNN and RF sentiment classification. The flow of proposed system is given in Figure 1.

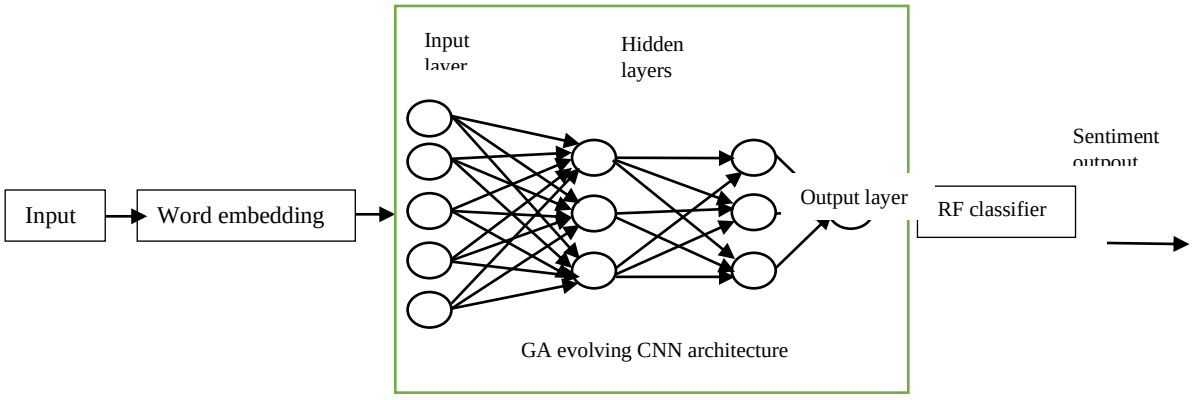


Figure 1. Proposed classification architecture

3.1. Word Embedding

The first step in the proposed system is to pre-process the text input. This step is necessary since the CNN cannot process the text input data. In this step the text input is converted into numerical values. The use of a word embedding model is a standard approach for generating vector values for text data. Word2Vec and GloVe are the two widely used word embedding models since they are proven to attain high performance. They are widely used in different text processing task and represent text as continuous vectors in low dimensional space. Word2Vector embedding is used in this work.

3.2. CNN architecture using GA for feature extraction]

In initial stage, CNN was only used for computer vision task and now it is capable to solve many natural language problems. CNN uses two dimensions for text inputs. One is used to represent the number of words and the other is the d-dimensional vector for each word. In this paper CNN is used for extracting important features for classification. The CNN has lots of parameters such as number of filters, filter size, number of hidden layers, epoch etc. These hyperparameters are the variables which determine the structure of CNN. The process of hyperparameter tuning in CNN will results in better performance. So genetic algorithm is used for performing hyperparameter tuning. The steps involved in designing CNN architecture using genetic algorithm is given in Algorithm 1.

There are two main requirements for genetic algorithms. First one is defining the solution domain and second is fitness function for every individual. The genetic algorithm work based on standard bio-inspired operations and will find best parameters for CNN network. The genetic approach has three phases: initialization, evaluation and update. These phases are explained detail in the following section.

3.2.1. Initialization phase

In the initial phase, the process is start by providing the vector input to CNN, size of the population, number of generations for the genetic algorithm and also provided different building blocks for CNN. There are X_{NXNP} solutions where N_x is the number of solutions and N_P is the number of parameters to be optimized. The hyperparameters used to control the CNN network configuration are initialization mode, epoch, dropout and learning rate. The values for these parameters are generated, then the crossover and mutations are set for the genetic parameters. The initial population phase is executed only once while the other GA phases are repeated. It is prominent phase in GA since it has a special role to improve GA performance. There are mainly two methods are used for population initialization in genetic algorithm. They are random initialization and heuristic initialization. The random initialization is one of the commonly used technique for generating initial seeds. The random initialization is used in the proposed model. The initial population is generated with completely random solutions. A random population of size N is generated by using the following equation

$$x_{i\,j} = l_j + random * (u_j - l_j)$$

Where

j ranges from 1 to N_p

iranges from 1 to N

$u_j$and $l_j$are upper and lower bounds for j^{th}parameter

For a particular solution $x_{i,}$ the possible solutions are given in Table 1. These values are pass to the evaluation phase.

3.2.2. Evaluation phase

The parameters selected from the initial phase is used to build the CNN model. The collected dataset is divided into training and testing sets based on 80:20 split method. For building CNN the training set is used and for evaluating the model on unknown data, the testing set is used.

3.2.3. Update phase

In this update phase, a best solution $x_b$which has high fitness value is selected. Then every solution in the current population is to be updated using the three operations of genetic algorithm such as mutation, crossover and selection. The evaluation and update phases are repeated until the termination condition is met.

Algorithm 1

Genetic algorithm for CNN architecture design

Input:

User review dataset D, Initial population X, Termination condition crossover and mutation probabilities

Output:

Set of most important features

- Given initial population X_{NXNP} where N is the number of solutions and N_p is the number of

Features used for designing CNN

- Divide the collected dataset into two groups such as training and testing set using the initial population
- For i=1, 2, ... I do the following

Evaluation:

- CNN_NET By using initial population X and training dataset, design the CNN architecture
- FIT_VAL (CNN_ NET) ← Calculate the fitness value for every CNN_NET by using the defined fitness function

Updation:

- $(CNN_NET)_x$Select the CNN_NET with best solution x

Crossover:

- Find the random split position and for any two possible solutions, two new solutions are created by swapping the information at the split point

Mutation:

- If there are non-crossover solutions, they should be muted

Selection

- Generating new population

End for

3.3. Random Forest Classifier

Random Forest proposed by Leo Breiman and Adele Cutler in 2001. This algorithm combines two concepts such as bagging and random subspaces. The random forest is a portion of family set methods which consider decision tree as an individual predictor. Random forest is one of the best among classification algorithms. It is capable to classify large amount of data with promising accuracy. It uses an ensemble method for classification problems and it generates number of decision trees during training phase. The classifier will output the class with major vote as the prediction result.

Algorithm2

Random Forest classification algorithm

Input

Set of features or variables

Output

Sentiment class

- For b= 1 to B do
  - Bootstrap sample Z* of size N from the training data should be drawn.
  - For each terminal node, iterate the following steps to generate the random forest tree T_b to the bootstrapped data
    - Select p variables at random from the q variables
    - C the best split-point/variable among the p
    - Split the current node in to two sub nodes
- Output the ensemble of trees $\{T_{b1}{}^B\}$
- The class of tree which have majority vote is considered as the classification output

TABLE 1. Sample parameter values

Parameters	Example value
Learning rate	0.001
Momentum	0.9
Iteration	8
Dropout	0.4

4. Dataset and experimental analysis

3.4. DATASET

3000 IMDB movie review dataset from Kaggle is collected for this work. These reviews have two fields such as review_text and sentiment. These labelled reviews can be any one of the five sentiment categories very negative, negative, neutral, positive and very positive. They are evaluated based on 5-point scale i.e.,

- Very negative (1)
- Negative(2)
- Neutral(3)
- Positive(4)
- Very positive (5)

The collected dataset is split into training and testing sets based on 80:20 split method.

3.5. EXPERIMENTAL RESULTS

The following base methods are used for comparing the performance of the proposed model.

Random Forest (RF):- The Traditional Random Forest classifier is built by using the vectorization methods such as TF-IDF and CountVectorizer. In this work the Random Forest Classifier is used as the base classifier since it is one of the ensemble method widely used in supervised machine learning. The accuracy obtained for the collected dataset using Random Forest Classifier is 86%.

Convolutional Neural Network (CNN): - A CNN network built with five convolutional layers, and max pooling operation. The word2Vec was the method used for converting the text input into vector from. This achieved an accuracy of 90%.

CNNRF:- A hybrid model called CNNRF also tested for the collected dataset. In this hybrid model CNN used to extract the features from the collected dataset, then the extracted features are given to the Random Forest classifier. The RF classifier performed the classification task and achieved an accuracy of 93%.

In order to improve the performance of the CNNRF proposed model, different hyperparameter tuning techniques is applied. The hyperparameters of the CNN which is tuned are learning rate, momentum, epoch, dropout value. Three different hyperparameter tuning techniques are used for experiment. They are

CNNRF(Random_Search):- Here, the hyperparameters of the CNN are tuned by using the random search method and the accuracy obtained is 94.5%.

CNNRF(Grid_Search):- In this model, the parameters of CNN are tuned by using the grid search method and it achieved 95.5% of accuracy

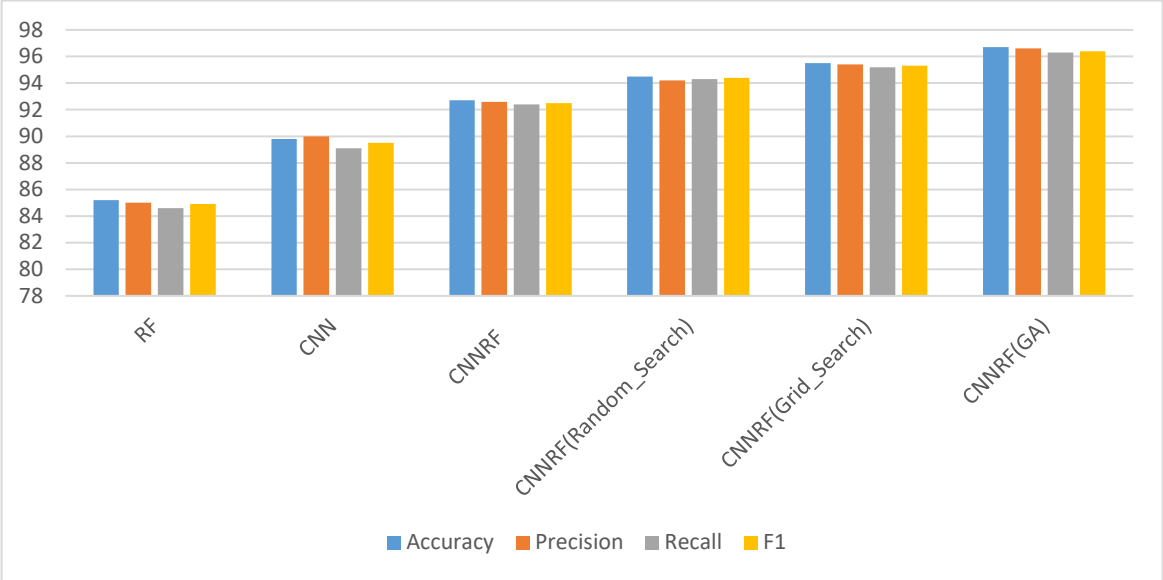
.

CNNRF(GA): -This is the proposed model. In this model, the CNN architecture is designed by using genetic algorithms. The performance of any neural network is determined by the value of the hyperparameters. In this proposed model, the CNN hyperparameters such as learning rate, momentum, epoch, dropout value is tuned and selected some best value for these parameters. This helped to find out best features for classification. Finally, the Random Forest classifier done the classification task and is given an accuracy of 96.3%.

The overall result of these experiments are given in Table 2.

Table 2. Experimental results

Fine-grained sentiment analysis				
Model	**Accuracy(%)**	**Precision(%)**	**Recall(%)**	**F1(%)**
RF	85.2	85	84.6	84.9
CNN	89.8	90	89.1	89.5
CNNRF	92.7	92.6	92.4	92.5
CNNRF(Random_Search)	94.5	94.2	94.3	94.4
CNNRF(Grid_Search)	95.5	95.4	95.2	95.3
CNNRF(GA)	96.7	96.6	96.3	96.4



The proposed method CNNRF with genetic algorithm hyperparameter tuning gave better accuracy as compared with other methods

The best CNN parameters obtained by using genetic algorithm optimization technique is given in Table 3. The accuracy on fail learning were tested by 0.01 to 0.02 and it is depicted in Figure 2.For learning rate 0.001, it gets maximum accuracy then it gets falls.

Table 3. Best Parameters

Parameters	Best Value
Learning rate	0.001
Momentum	0.95
Iteration	7
Dropout	0.4

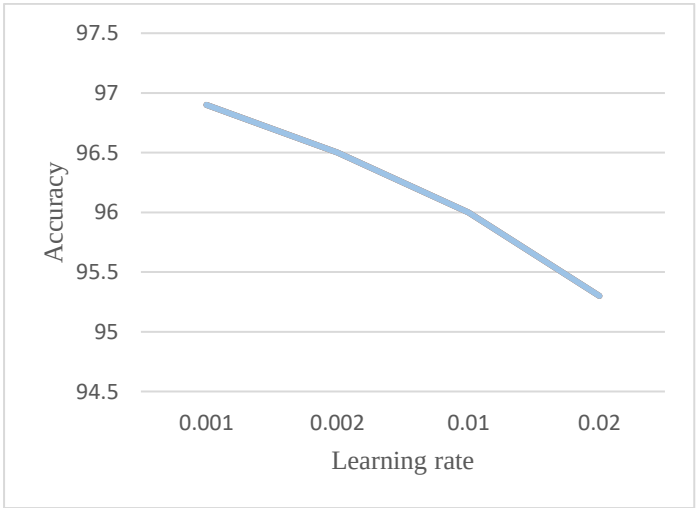


Figure 2. Accuracy on fail learning rate by 0.01 and 0.02

The hyperparameter tuning of the proposed model using genetic algorithm helped to identify the best values for the selected parameters and provided promising performance as compared with grid search and random search tuning methods.

5. Conclusion

In this paper, a genetic algorithm-based CNN architecture with Random Forest classifier is proposed for fine grained sentiment prediction. For this purpose, learning rate, momentum, iteration and dropout values of CNN architecture are tuned. Hyperparameter tuning using genetic algorithm, grid search and random search performed

on collected dataset. Most important features are extracted using CNN and the classification on these selected features are performed by Random Forest classifier. The result shows that the genetic algorithm based hyperparameter tuning method provided promising accuracy as compared with the other two tuning methods.

References

1. Neha Bansal,Arun Sharma, R. K. Singh, "An Evolving Hybrid Deep Learning Framework for legal document classification", Ingénierie des Systèmesd'Information, Vol.24, pp:425- 431, 2019.
2. AshrayBhandare and Devinder Kaur," Designing Conventional Neural Network Using Genetic Algorithms", Int'l Conf. Artificial Intelligence,ICAI'18, 2018.
3. Xuell Xiao, Ming Yan, Sunitha Basodl, Chunyan Ji and Yi Pan, "Efficient Hyperparameter Optimization in Deep Learning Using a Variable Length Genetic Algorithm",
4. Petro Liashchynskyi and PavloLiashchynskyi, "Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS", arXiv, CS-LG, Dec 2019.
5. Adrian Catalin and Razvan Andonie, "Weighted Random Search for Hyperparameter Optimization", International Journal of Computers Communications and Control", Vol. 14, 2019.
6. Haiman Tian, Samira Pouyanfar, Jonathan Chen, Shu-Ching Chen and Sitharma S.," Automatic Conventional Neural Network for Image Classification Using Genetic Algorithms", IEEE International Conference on Information Reuse and Integration (IRI), 2018.
7. Yanan Sun, Bing Xue, Mengjie Zhang, Gary G. Yen and JianchengLv, "Automatically Designing CNN Architecture Using Genetic Algorithm for Image Classification", Neural and Evolutionary Computing, Jan 2019.
8. Amr AbdelFatah Ahmed, Saad M., Mohamed M., "A Novel Automatic CNN Architecture Approach Based on Genetic Algorithm", International Conference on Advanced Intelligent Systems and Informatics, Oct 2019.
9. Mohammad Ehsan, MoloudAbdar, Mehmet Akif, Shahla Nemati, U.Rajendra Acharya, " A novel method for sentiment classification of drug reviews using fusion of deep and machine learning techniques", Knowledge Based Systems", Vol. 198, June 2020.
10. DOI: https://doi.org/10.1016/j.knosys.2020.105949.
11. AlperKursatUysal, Yi Lu Murphey, "Sentiment Classification: Feature Selection Based Approaches Versus Deep Learning", International Conference Computer and Information Technology (CIT), 2017.
12. MahaHeikal, MarvanTorki, Nagwa El-Makky, "Sentiment Analysis of Arabic Tweets using Deep Learning", Procedia Computer Science, Elsevier, 2018.
13. Yi Yang, Ying Li, Jin Wang, R. Simon Sherratt, "Sentiment Analysis for E-commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning", IEEE Access, vol. 8, 2020.
14. Akshi Kumar, Kathiravan Srinivasan, Cheng Wen-Huang, Albert Y., "Hybrid Context Enriched Deep Learning Model for Fine-grained Sentiment Analysis in Textual and Visual Semiotic Modality Social Data", Information Processing and Management, Elsevier, 2020.
15. Oscar Araque, Ignacio, J.Fernando, Carlos A., "Enhancing Deep Learning Sentiment Analysis with Ensemble Techniques in Social Applications", Expert Systems With Applications, Elsevier, 2017.
16. Zhaoxia wang, Chee Seng, Yinping Yang, Seng Beng," Fine-grained sentiment analysis of social media with emotion sensing", Future Technologies Conference, Singapore, 2016.
17. Ramesh Wadawadagi, Veerappa Pagi, "Sentiment Analysis with Deep Neural Networks: comparative study and performance assessment", Artificial Intelligence Review, Springer, 2020.
18. Shahnoor C. Eshan, Mohammad S Hasan, "An Application of Machine learning to Detect Abusive Bengali Text", 20th International Conference of Computer and Information Technology(ICCIT), December 2017.
19. Muhammad Murtadha Ramadhan, ImasSukaesihSitanggaang, FahrediRizkyNasution and Abdullah Ghifari," Parameter Tuning in random Forest Based on Grid Search Method for Gender Classification Based on Voice Frequency", International Conference on Computer, Electronics and Communication Engineering, 2017.
20. Xingzhi Zhang, Yan Yang, Zhurong Zhou," A Novel Credit Scoring Model based on Optimized Random Forest*",* 8th Annual Computing and Communication Workshop and Conference (CCWC*)*, 2018.
21. Hitesh H Parmar, Sanjay Bhanderi, Glory Shah," Sentiment Mining of Movie Reviews using Random Forest with Tuned Hyper-parameters",International Conference on Information Science, July 2014.
22. R. Xia, C. Zonga, S. Li," Ensemble of feature sets and classification algorithms for sentiment classification", Information Sciences, Elsevier, vol. 181, pp.1138-1152, 2011.
23. Yashaswini Hegde, S.K. Padma," Sentiment Analysis Using Random Forest Ensemble for Mobile Product Reviews in Kannada", 7th International Advance Computing Conference(IACC), *IEEE*, 2017.
24. BagusSetyaRintyarna, RiyanartoSarno, ChastineFatichah, "Enhancing the Performance of Sentiment Analysis Task on Product Reviews by Handling Both Local and Global Context", International Journal of Information and Decision Sciences*",* February 2020.

25. M. Ahmad, S. Aftab, S.S Muhammad, and S. Ahmad," Machine learning techniques for sentiment analysis- A review", International Journal of Multidisciplinary Science and Engineering, vol. 8, no. 3, pp. 27-32, 2017.
26. S. Zhang, Z. Wei, Y. Wang, and T. Liao," Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary", Future Generation Computer Systems, vol. 81, pp. 395-403, April 2018.