Ensemble-based method designed to Increase Sentiment Analysis Categorization Accuracy.

[1]Vaishali Malik, [2]Dr. Nidhi Tyagi

[1]Research Scholar, Shobhit Institute of Engineering & Technology, Meerut (U.P.),
vaishalimalik07@gmail.com
[2]Professor, Shobhit Institute of Engineering & Technology, Meerut (U.P.),
nidhi.tyagi@shobhituniversity.ac.in

KEYWORDS
Ensemble learning, accuracy, datasets, Naive Bayes, MaxEnt.

ABSTRACT:
The need for understanding user behaviour is strong due to the ever-increasing volume of data generated by social media users, especially in light of the current coronavirus outbreak. In this experiment, we focus on a dataset that includes the subjective assessments of those who have written about the epidemic. It is not easy to find the best classification methods for this sort of data. When compared to traditional feature-based methods, deep learning models for sentiment analysis may provide more nuanced representations and better overall performance. The suggested method is an ensemble strategy that makes use of boosting in addition to supervised machine learning methodologies. The goal is to examine the usefulness of the idea of using several classifier systems on IMDb and other multi-domain datasets. We have seen implementations of the Vote method in combination with Naive Bayes, Maximum Entropy, and Boosting classifiers. The Ensemble method outperforms both the best-reported individual classifier (Support vector machines) and Naive Bayes (which is widely used). Precision, recall, and accuracy are only few of the metrics that are used to evaluate the effectiveness of various approaches.

INTRODUCTION

A large number of evaluations were collected because of the rising popularity of social media and advertising platforms online. To inform the suggested method for making qualitative judgements, we will gather and aggregate the reviews. Since the reviews are often unstructured, they need processing such as categorization or clustering in order to provide useful results. The current method for classifying text reviews makes use of TF-IDF, n-gram, and supervised machine learning techniques. Naive Bayes (NB), Maximum Entropy (ME), Stochastic Gradient Descent (SGD), and Support Vector Machine (SVM) are only few of the machine learning techniques that have been investigated for classifying human emotions. Consequently, the SVM classifier may provide results that are more accurate.

Machine learning methods based on supervised and unsupervised learning are the two most common approaches to sentiment analysis. If the dataset is labelled, then supervised learning may be used to train for plausible results that aid in decision-making. Since unlabelled data are not required for the unsupervised learning process, processing them is more difficult. Clustering methods are used to address the challenge of processing unlabelled data.
We see sentiment analysis being performed on three distinct scales: the document level, the phrase level, and the aspect level. An opinion is classified as favourable, negative, or neutral at the document level. In order to assess if a statement is negative, positive, or neutral, we look at the words that make up the phrase. At the aspect level, analysis focuses on all emotional expressions in the text and the aspect to which they pertain. Figure 1 depicts the whole procedure for sentiment analysis.

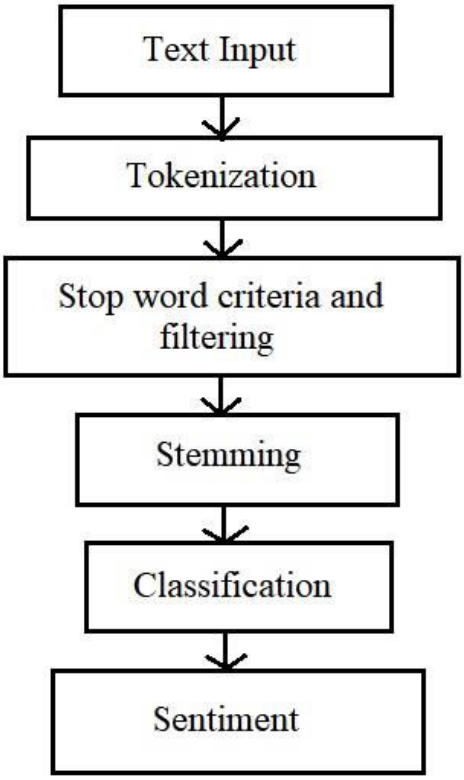

Figure 1: Process of sentiment analysis

The process of categorising a collection involves placing its constituents into predetermined groups. To successfully classify data, one must reliably anticipate which of many target classes best applies to each individual example. It is common practise to separate the categorization data into training and testing sets when working with a machine-learning algorithm. An appropriate classification algorithm will discover connections between the predictor and target values. Relationship-finding strategies vary between categorization algorithms. A model is constructed to summarise these connections; it may then be used on another data set for which the category labels are unknown.

Machine learning approaches in sentiment analysis

The machine learning strategy treats sentiment analysis as a standard text classification issue, and uses well-known ML algorithms to find solutions based on syntactic and/or linguistic factors. By removing unimportant and noisy characteristics, feature selection helps machine learning algorithms work with a smaller, more manageable feature space.

Supervised Learning

The process of inferring a function from labelled data is known as supervised learning in machine learning. A collection of training examples makes up the training data. Each example in supervised learning consists of a vector of input objects and an expected value for those vectors (also called as the supervisory signal). An inferred function is generated from the analysis of the training data and may be used to the mapping of untrained instances. Regression problems and classification issues are two subsets of supervised learning. The classification process in data mining involves sorting data into predetermined groups. To successfully classify data, one must reliably anticipate which of many target classes best applies to each individual example. Predicting numbers is the job of the regression data mining function. Regression is useful for determining how a variable responds to changes in other variables. Predicting and predicting events is a common use case for regression analysis.

Unsupervised Learning

Using unlabelled data, unsupervised learning is a machine learning approach that looks for patterns. In the realm of unsupervised learning, issues may be broken down further into two categories: clustering and association. Clustering is the process of dividing a dataset into smaller groups (clusters) based on their shared characteristics. Similarity metrics, such as Euclidean or probabilistic distance, are used to determine the clusters used in the modelling process. Association is a data mining technique used to discover potential rules that regulate causal relationships and links between groups of data.

Ensemble Learning

To reduce variance (by bagging), bias (by boosting), or enhance predictions, ensemble methods are meta-algorithms that integrate many machine learning approaches into a single predictive model (stacking). There are two broad categories of ensemble techniques: sequential and parallel. Sequential ensemble approaches produce their initial base learners one at a time (e.g. AdaBoost). The driving force for sequential techniques is the need to take advantage of the interdependence among the learners at the foundation. Increasing the weight of previously mislabelled cases may improve overall performance.

BACKGROUND AND RELATED WORK

The basic learner in parallel ensemble techniques are produced simultaneously (e.g. Random Forest). Since errors may be drastically minimised by averaging across several base learners, exploiting their independence is the driving force for parallel techniques (Yaeger 2010; Zhang & Wu 2012) [9]. Sentiment classification of twitter data for airline services was performed by Wan and Gao (2015) [10] using an ensemble sentiment classification strategy based on the Majority Vote principle of multiple classification methods. These methods included Naive Bayes, SVM, Bayesian Network, C4.5 Decision Tree, and Random Forest. The findings on this Twitter dataset from an airline service demonstrate that the suggested ensemble strategy is superior than individual classifiers (Manek et al. 2017) [11].

Catal and Nangir (2017) [12] introduced a unique approach to sentiment classification for the Turkish sentiment classification issue, based on Vote ensemble classifier that integrates features from three separate classifiers (Bagging, Naive Bayes, and Support Vector Machines). The proposed method outperformed both Naive Bayes and Support Vector Machines, the best reported individual classifiers for the datasets used.

Examining the author's perspective on a topic or the document's general contextual polarity is what sentiment analysis is all about [13]. Sentiment analysis aims to categorise writings not by content but by how they make the reader feel [14]. In order to identify and extract subjective information from a huge amount of unstructured data, sentiment analysis makes use of information retrieval, natural language processing, data mining, and knowledge management methods [15]. According to [16], assessing the tone of the source materials by sentiment analysis is a multi-step procedure that entails five distinct steps. A few examples of these steps include data collection, text preparation, sentiment detection, sentiment categorization, and results display. Both supervised learning and unsupervised learning are commonly used implementations of the sentiment analysis technology [17]. Sorting the training set to generate textual patterns is an integral part of the supervised learning methodology. To train its models, unsupervised learning relies not on a pre-existing database but rather on a predetermined collection of words, that includes both positive and negative examples. As a result, the document may be tagged using the negative and positive phrases most often found in the text [18, 19].

Several different industries make use of sentiment analysis. According to [20]'s writers, analysing public opinion on social media platforms using sentiment analysis helps the government figure out where it excels and where it needs improvement. Similarly, in e-commerce, sentiment analysis is used to change negative reviews about product quality into positive ones [21]. Sentiment analysis, as confirmed by Vohra & Teraiya [22], is used to evaluate the feedback provided by consumers on a company's goods or services. An excellent example of a real-time Twitter analysis programme is

Tweetfeel [23]. Similarly, Wang et al. [24] emphasised the use of sentiment analysis in Blogger-centric contextual advertising, which entails tailoring adverts on blog sites to the specific preferences of individual businesses. In light of these results, sentiment analysis has been extensively used in a variety of settings for the purpose of detecting and evaluating certain patterns of behaviour and attitude [25].

PROPOSED APPROACH

There is processing done on the IMDb and Multi-domain dataset reviews to get rid of stop words and other irrelevant data. Afterwards, vectorization methods are used to convert the textual information into a numerical matrix. The suggested ensemble method is shown in Figure 2. Adaboosting is a method that improves SVM's efficiency. For classification purposes, the vote technique has been combined with Naive Bayes, Support Vector Machine (SVM), and Maximum Entropy.

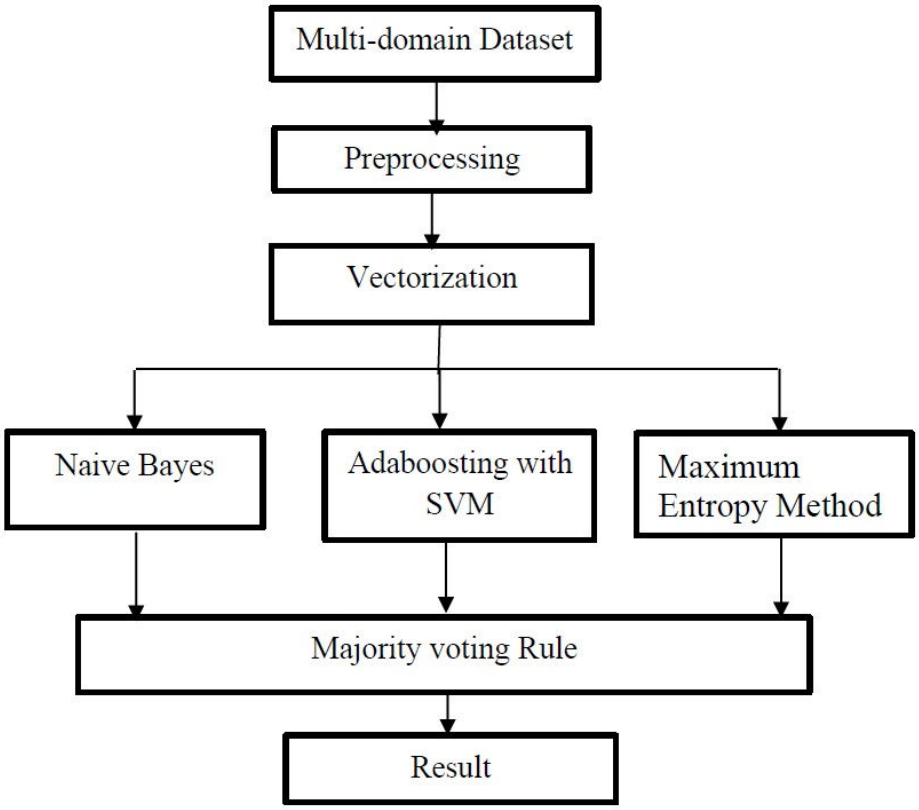

Figure 2: Proposed flow diagram of used approach

Data Pre-processing, Vectorization and Count Vectorization

Document sets from different domains, such as those containing movie reviews, are categorised according to the polarity of the emotions they express. We take into account the Internet Movie Database (IMDb) dataset, which contains 12,500 favourably labelled reviews and 12,500 adversely labelled reviews. Reviews from many different fields, such as literature, cooking, movies, and consumer electronics, are included in the multi-domain dataset. There are a total of 3,000 reviews across all domains, with 1,000 rated as "good," 1,000 rated as "bad," and 1,000 left unrated. It is common practise in the field of data mining to first pre-process raw data into a more manageable format. Sometimes even the text reviews include completely illogical information that must be filtered out before categorization.

Three distinct supervised machine learning techniques are then used to categorise the numerical matrix resulting from the text reviews.

Algorithm 1: Ensemble Algorithm

Data Input: [Reviews of movies, books, DVDs, gadgets, and home appliances]
Data Preprocessing
    For each Data in **Data_Input**
    Remove Stop words, Numerical characters
    Calculate TF-IDF
End
Learner: [Naive Bayes, MaxEnt, SVM, Boosting, Ensemble technique]
    For each Data
    TRAIN = random 80% of data
    TEST = data – TRAIN
    PREDICTOR = Train LEARNER with TRAIN
    [Precision] = PREDICTOR on TEST
    [Recall] = PREDICTOR on TEST
    [Accuracy] = PREDICTOR on TEST
End

RESULTS AND DISCUSSION

Parameters from a matrix called a confusion matrix or contingency table are used to assess the effectiveness of a supervised machine learning system. It is used to evaluate the efficacy of an algorithm in a supervised machine learning setting. For this study, we compiled collections of reviews across many domains, including those for movies.

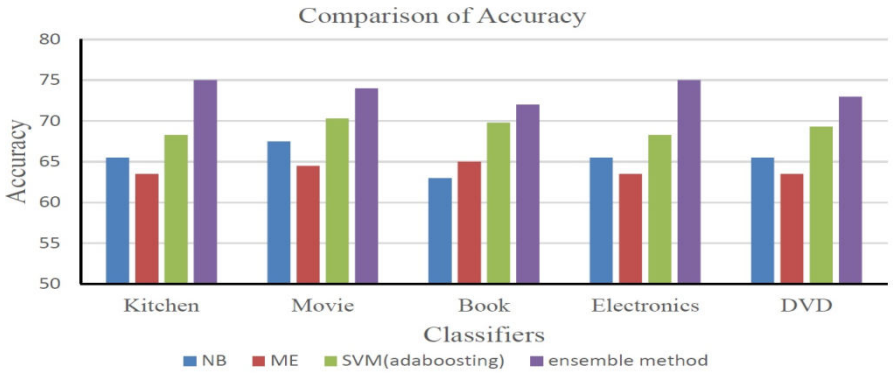

Figure 3: Accuracy of Multi-domain dataset

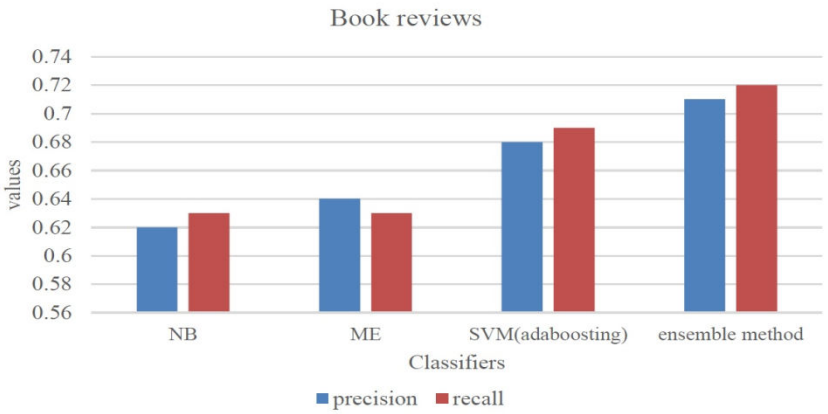

Figure 4: Precision and Recall of book reviews

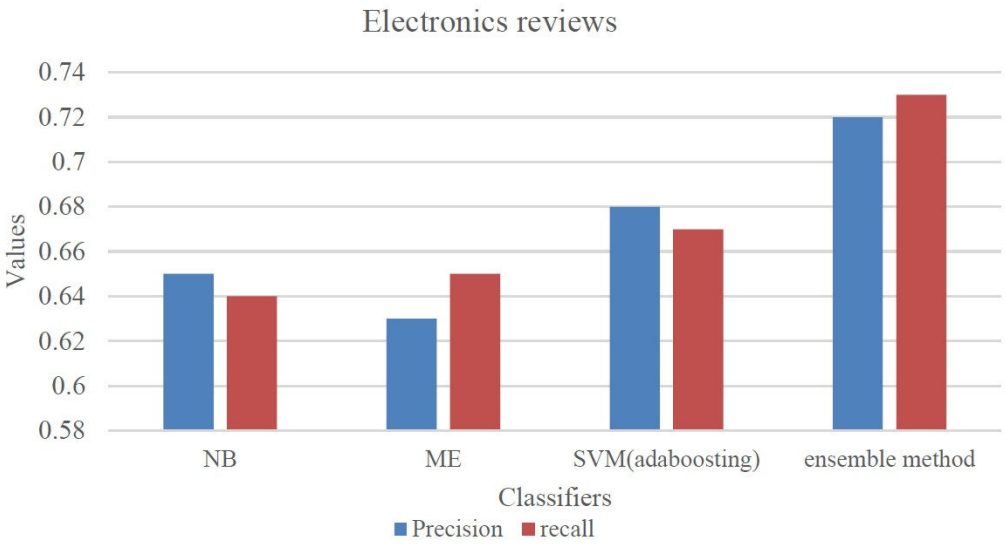

Figure 5: Precision and Recall of electronics reviews

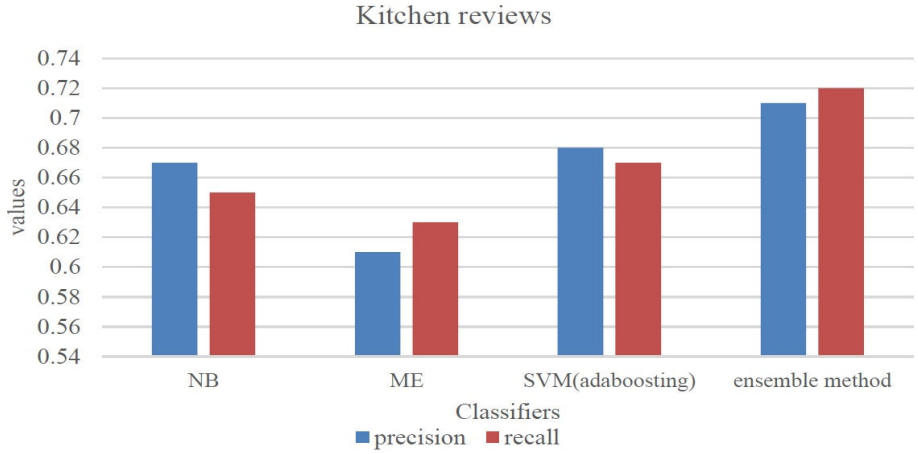

Figure 6: Precision and Recall of kitchen reviews

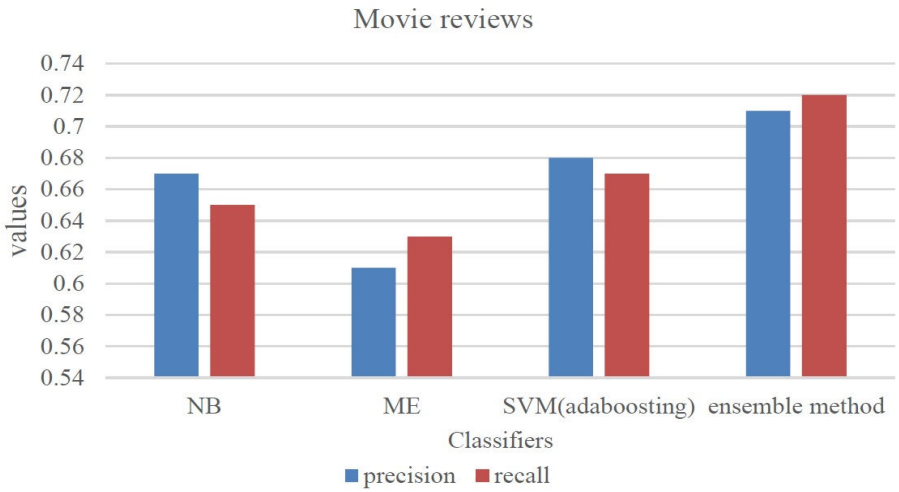

Figure 7: Precision and Recall of movie reviews

There are a total of 4,500 processed reviews in the dataset, 2,000 of which are favourable and 2,500 of which are negative. The obtained result demonstrates that the suggested method outperforms the alternatives in the machine-learning field. Figures 4, 5, and 6 provide a comparison of the precision and recall graphs for multi-domain and other stated reviews.

CONCLUSION

The n-gram technique improved accuracy, but unigram characteristics revealed very modest differences, therefore the ensemble approach was explored with a unigram approach to minimise complexity. Using supervised machine learning methods, the suggested solution classifies IMDb and multi-domain datasets. The majority voting rule used algorithm categorization findings. Majority vote determined classification accuracy. This strategy improves accuracy over individual machine learning algorithms. Combining Nave Bayes, Maximum Entropy, and Adaboosting (SVM) improves accuracy. Different approaches are evaluated for precision, recall, and accuracy. This technique classifies evaluations as positive or negative, so it may advise individuals by alerting them of the product's shortcomings or strong qualities they need to integrate to maintain market competitiveness. The ensemble technique improves accuracy by 5% to 10% for kitchen and electronics datasets and 3% to 7% for other datasets.

REFERENCES

[1] Reddy, G. Thippa, Sweta Bhattacharya, S. Siva Ramakrishnan, Chiranji Lal Chowdhary, Saqib Hakak, Rajesh Kaluri, and M. Praveen Kumar Reddy. "An ensemble based machine learning model for diabetic retinopathy classification." In *2020 international conference on emerging trends in information technology and engineering (ic-ETITE)*, pp. 1-6. IEEE, 2020.

[2] Kadavi, Prima Riza, Chang-Wook Lee, and Saro Lee. "Application of ensemble-based machine learning models to landslide susceptibility mapping." *Remote Sensing* 10, no. 8 (2018): 1252.

[3] Jonsson, Leif, Markus Borg, David Broman, Kristian Sandahl, Sigrid Eldh, and Per Runeson. "Automated bug assignment: Ensemble-based machine learning in large scale industrial contexts." *Empirical Software Engineering* 21, no. 4 (2016): 1533-1578.

[4] Zhang, Cha, and Yunqian Ma, eds. *Ensemble machine learning: methods and applications*. Springer Science & Business Media, 2012.

[5] Dong, Xibin, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. "A survey on ensemble learning." *Frontiers of Computer Science* 14, no. 2 (2020): 241-258.

[6] Liu, Nan, and Han Wang. "Ensemble based extreme learning machine." *IEEE Signal Processing Letters* 17, no. 8 (2010): 754-757.

[7] Zubair Hasan, K. M., and Zahid Hasan. "Performance evaluation of ensemble-based machine learning techniques for prediction of chronic kidney disease." In *Emerging research in computing, information, communication and applications*, pp. 415-426. Springer, Singapore, 2019.

[8] Behera, Rabi Narayan, Manan Roy, and Sujata Dash. "Ensemble based hybrid machine learning approach for sentiment classification-a review." *International Journal of Computer Applications* 146, no. 6 (2016): 31-36.

[9] Troussas, Christos, Akrivi Krouska, and Maria Virvou. "Evaluation of ensemble-based sentiment classifiers for Twitter data." In *2016 7th international conference on information, intelligence, systems & applications (IISA)*, pp. 1-6. IEEE, 2016.

[10] Subba, Basant, and Simpy Kumari. "A heterogeneous stacking ensemble based sentiment analysis framework using multiple word embeddings." *Computational Intelligence* 38, no. 2 (2022): 530-559.

[11] Dashtipour, Kia, Cosimo Ieracitano, Francesco Carlo Morabito, Ali Raza, and Amir Hussain. "An ensemble based classification approach for persian sentiment analysis." In *Progresses in Artificial Intelligence and Neural Systems*, pp. 207-215. Springer, Singapore, 2021.

[12] Behera, Rabi Narayan, Manan Roy, and Sujata Dash. "Ensemble based hybrid machine learning approach for sentiment classification-a review." *International Journal of Computer Applications* 146, no. 6 (2016): 31-36.
[13] Akhtar, Md Shad, Deepak Gupta, Asif Ekbal, and Pushpak Bhattacharyya. "Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis." *Knowledge-Based Systems* 125 (2017): 116-135.
[14] Kanakaraj, Monisha, and Ram Mohana Reddy Guddeti. "NLP based sentiment analysis on Twitter data using ensemble classifiers." In *2015 3Rd international conference on signal processing, communication and networking (ICSCN)*, pp. 1-5. IEEE, 2015.
[15] Tian, Yuanhe, Guimin Chen, and Yan Song. "Aspect-based sentiment analysis with type-aware graph convolutional networks and layer ensemble." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2910-2922. 2021.
[16] Mohammadi, Azadeh, and Anis Shaverizade. "Ensemble deep learning for aspect-based sentiment analysis." *International Journal of Nonlinear Analysis and Applications* 12, no. Special Issue (2021): 29-38.
[17] Ibrahim, Ahmed. "Forecasting the early market movement in bitcoin using twitter's sentiment analysis: An ensemble-based prediction model." In *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pp. 1-5. IEEE, 2021.
[18] Gopalakrishnan, Vinodhini, and Chandrasekaran Ramaswamy. "Sentiment learning from imbalanced dataset: An ensemble based method." *International Journal of Artificial Intelligence* 12, no. 2 (2014): 75-87.
[19] Yadav, Konark, Aashish Lamba, Dhruv Gupta, Ansh Gupta, Purnendu Karmakar, and Sandeep Saini. "Bi-LSTM and ensemble based bilingual sentiment analysis for a code-mixed Hindi-English social media Text." In *2020 IEEE 17th India Council International Conference (INDICON)*, pp. 1-6. IEEE, 2020.
[20] Vinodhini, G., and R. M. Chandrasekaran. "A sampling based sentiment mining approach for e-commerce applications." *Information Processing & Management* 53, no. 1 (2017): 223-236.
[21] Wang, Gang, Jianshan Sun, Jian Ma, Kaiquan Xu, and Jibao Gu. "Sentiment classification: The contribution of ensemble learning." *Decision support systems* 57 (2014): 77-93.
[22] Moung, Ervin Gubin, Chai Chuan Wooi, Maisarah Mohd Sufian, Chin Kim On, and Jamal Ahmad Dargham. "Ensemble-based face expression recognition approach for image sentiment analysis." *International Journal of Electrical & Computer Engineering (2088-8708)* 12, no. 3 (2022).
[23] Sultana, Naznin, and Mohammad Mohaiminul Islam. "Meta classifier-based ensemble learning for sentiment classification." In *Proceedings of International Joint Conference on Computational Intelligence*, pp. 73-84. Springer, Singapore, 2020.
[24] Vinodhini, G., and R. M. Chandrasekaran. "A comparative performance evaluation of neural network based approach for sentiment classification of online reviews." *Journal of King Saud University-Computer and Information Sciences* 28, no. 1 (2016): 2-12.
[25] Sunitha, D., Raj Kumar Patra, N. V. Babu, A. Suresh, and Suresh Chand Gupta. "Twitter sentiment analysis using ensemble based deep learning model towards COVID-19 in India and European countries." *Pattern Recognition Letters* 158 (2022): 164-170.