A Comparative Analysis of Different Machine Learning Classification Models for Sentiment Analysis

Aryaman Jain[1] · Vanshita Tongia[1]

[1]Medi-Caps University, Indore, India

Abstract:**With the world transforming into a digital age, the generation of textual documents is increasing at an unprecedented rate. This has consequently given rise to the need to organize these documents into proper categories and structure. Text classification, also known as text categorization, is the process of categorizing text into organized groups. In this paper, IMDB dataset of fifty thousand movie reviews is assessed and a classification system is designed. It compares Linear SVC, Bernoulli Naive Bayes, Logistic Regression, Multinomial Naïve Bayes and Random Forest as classification algorithms for applying sentiment analysis and finding the polarity of the given review. These classifiers were tested, analysed and compared with each other and finally a conclusion was obtained.The authors decided to show the comparison based on several parameters such as precision, accuracy, F1-score, recall and confusion matrix. The classifier which gets the highest among all these parameters is termed as the best machine learning algorithm for the text sentiment analysis of IMDB review data set.**

Index Terms **-Text Classification · Sentiment Analysis · Machine Learning · Logistic Regression · Random Forest · Multinomial Naïve Bayes · Bernoulli's Naïve Bayes · Linear Support Vector Classifier · Natural language processing**

I. INTRODUCTION:

Improving sales and retaining customers are the fundamental goals of a business. For a business to thrive, it must understand the customer's response to a particular service or product. In this "data age", a huge amount of data is generated every day. A plethora of consumer opinions are easily and readily available on various online sources in the form of blogs, forums, social media, review sites and more, which act as a base for many people to make their decisions on. But it is beyond the control of manual techniques to analyse millions of reviews and aggregate them into a quick and efficient decision. To overcome this problem, machine learning plays an important role. Machine learning is a subset of artificial intelligence (AI) that focuses on using statistical techniques to create intelligent computer systems that learn from existing databases. It is used in various fields such as image processing [1], speech recognition[2], weather forecasting [3], and others.Sentiment analysis, also known as opinion mining, is a machine learning tool that uses computing power to understand the underlying emotions in text. It analyses the text by finding out the semantic meaning of the text and thus explains the polarity of the text, from positive to negative. By training machine learning tools with examples of emotions in text, machines automatically learn to detect emotions without human intervention. By categorizing content into different categories, users can easily find anything. After the text is grouped, it is analysed by various models. This model has the task of applying tags to content. These models are just machine learning algorithms, also known as classifiers. These classifiers need to be trained to make predictions on the text dataset. These classifiers are trained by assigning the tag and then associating it with text fragments.

II. RELATED STUDY:

- **RELATED STUDY ON LOGISTIC REGRESSION**

  Prabhat and Khullar [4] have presented that the vast amount of online data cannot be mined constructively to extract valuable information and be a credible base for decision making. Sentiment analysis is a method in which we judge people's ideas, thoughts, opinions and belief about a particular idea or entity. Authors performed sentiment classification on data using Naive Bayes classifier and logistic regression. They used supervised and unsupervised learning algorithms. Criterions for the effectiveness of algorithms were based on accuracy, precision and throughput. The analysis using logistic regression came out to have 10.1% more accuracy and 4.34% more precision with approximately one-fifth implementation time for same size of data set compared to Naive Bayes classifier.

- **RELATED STUDY ON LINEARSVC**

  Support Vector Machines (SVM) is a powerful; state-of-the art algorithm with strong theoretical foundations based on the Vapnik-Chervonenkis theory [5].

  SVM constructs hyper-planes in a multidimensional space that separates different class boundaries and the number of dimensions is called the feature vector of the dataset.

- **RELATED STUDY ON BERNOULLI'S NB:**

  Naive Bayes is a machine learning classification algorithm based on Bayes' theorem that gives the probability of an event occurring. The Naive Bayes classifier is a classifier which focuses on probability. That is, given an input, it predicts the probability that the input will fall into all classes. Also called conditional probability.Bernoulli Naive Bayes is a type of Naïve Bayes. Discrete data is used in this and it works with Bernoulli distribution. The main feature of Bernoulli Naive Bayes is that it only accepts features as binary values such as true or false, yes or no, success or failure, 0 or1.

$$P(x_i \mid y) = P(i \mid y)x_i + (1 - P(i \mid y))(1 - x_i)$$

P(A|B) – the probability of event A occurring, given event B has occurred
P(B|A) – the probability of event B occurring, given event A has occurred
P(A) – the probability of event A
P(B) – the probability of event B

- **RELATED STUDY ON RANDOM FOREST:**
  Random forest is one of the most potent ensemble methods with great performance when it comes to high-dimension data, according to Nadi and Moradi [6]. By increasing the number of trees and decreasing the number of levels for each tree in random forest, the authors of this paper have presented a new approach to improve random forest performance. In this method, the trees are confined to a specific depth to allow for greater views. Every tree that is bound is regarded as a local view of the issue, and the more local the view, the better the classification. The outcomes demonstrated that the accuracy for high-dimension problems can be improved by binding the trees.

III. METHODOLOGY:

There are many machine learning algorithms used for text classification. But not all have the same accuracy or precision. Some are less accurate; some are more accurate. In this paper, text classification is done using five different machine learning algorithms implemented on the IMDB movie ratings dataset. The classification algorithms used are Logistic Regression Classifier, Random Forest Classifier, Bernoulli Naive Bayes Classifier, Multinomial Naive Bayes Classifier, and Linear Support Vector Classifier. They reported considerable accuracy and demonstrated efficient processing of dataset. Each of these five algorithms works in a completely different way. One works with specific formulas for classification and prediction, the other works by building nodes and trees (random forest). Each of these algorithms has developed a solution for classifying text.
The overview of architecture is shown in fig. 1. At the beginning, we import the necessary libraries which can also simultaneously be included in code as we proceed further. Next, we load the data set on which we need to perform the classification. To evaluate the representation, we used the IMDB movie review data set. The data set contains two columns (i) review of the movie and (ii) sentiment, which is categorized as either positive or negative. Then, we perform the pre- treatment or cleaning of text. After this step, the focus comes on the representation of text. In the testing phase, the five classifiers are applied on the data set which gave five parameters as the output and also to compute which feature has the maximum value for a particular class in the dataset [7, 8]. The parameters computed are accuracy, precision, F1-score and confusion matrix.

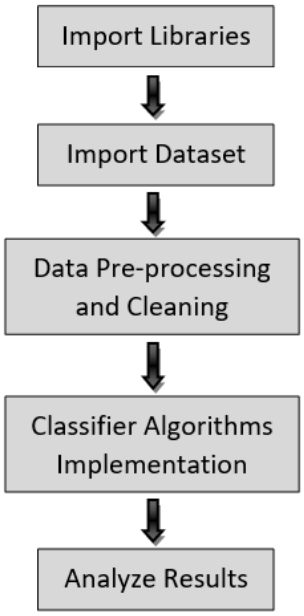


Fig. 1 Architecture of the implementation

TEXT PRE-TREATMENT AND REPRESENTATION:

We start by removing noisy text, which is basically html strips, brackets and special characters, Then, we perform text stemming, which is a kind of text normalization. The sentence can be written in different ways by changing the tense, but the meaning is the same. So, Stemmer helps you get rid of these tenses by making sentences have the same meaning. Stemmer's main task is to summarize sentences. The algorithm used for stemming is Porter Stemmer, which performs condensation tasks. Moving on, there are some words that do nothing to distinguish classes. These words can be prepositions, conjunctions, and pronouns. These words have no context in the text because they do not contribute to classification [9].

These words are called stop words. Therefore, stop words such as "that", "a", "and", "but", "or" should be removed. Therefore, English stop words were downloaded. After downloading the stop word list, we need to compare the words against this list and filter the words from the list [10]. After pre-processing and cleaning, we obtained the output of cleaned text as shown in Table 1.
Then we categorized our reviews on the basis of their lengths. We made six different groups with each group containing review of different lengths. The groups formed were (i) less than 50 words (ii) 50 to 100 words and so on until group containing reviews of length 250 to 300 words. Reviews containing more than 300 words were discarded. The final size of the groups can be seen in Fig. 2.

S. No.	Original Review	Cleaned Text
1	One of the other reviewers has mentioned that ...	one review ha mention watch 1 oz episod youll ...
2	Basically there's a family where a little boy ...	basic famili littl boy jake think zombi hi clo...
3	I thought this was a wonderful way to spend ti..	thought thi wa wonder way spend time hot summe...
4	Petter Mattei's "Love in the Time of Money" is...	petter mattei love time money visual stun film...

Table 1- Cleaning the text Before v/s After

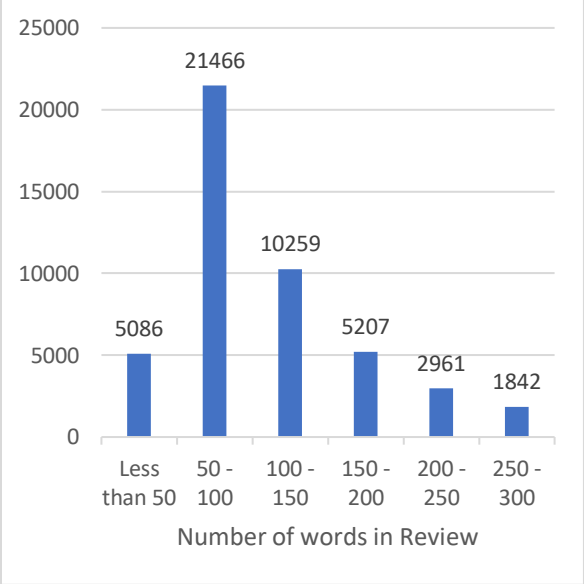


Fig. 2Number of reviews in each category

IMPLEMENTATION OF CLASSIFIER:
After performing text pre-treatment and representation, the classifiers are now to be implemented. We had considered five classifiers to determine which has the best output. In the beginning, we had split the data set into training and testing set, the size of the testing set being 30% and training set being 70%. As all the classifiers implemented are supervised so the data sets are provided and one has to just apply classification algorithms.

IV. RESULTS AND OUTCOME:

After running the code successfully, the required output was obtained. As mentioned earlier, the comparison between the algorithms is done on the basis of five parameters namely accuracy, precision, F1-score and confusion matrix. Let us compare the five algorithms implemented one by one.

Confusion matrix, as seen in fig 3, is defined as how much a classifier is able to predict the correct value, i.e., true positives in classification or what number of values belongs to the correct class rather than the other class.

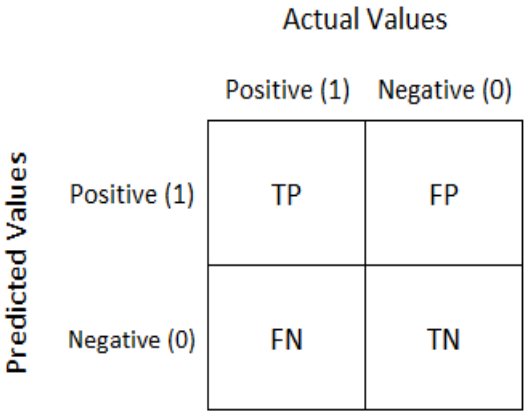


Fig. 3Confusion Matrix

Accuracy is defined as the ratio of observations predicted correctly to the total number of observations. The mathematical formula is defined as:

Accuracy = TP + TN / (TP + TN + FP + FN)

Precision is defined as the ratio of observations that are predicted positively correct to the total number of observations predicted positively. The mathematical formula is defined as:

Precision = TP / TP + FP

Recall is defined as the ratio of observations that are predicted positively correct to the total number of observations in an actual class. The mathematical formula is defined as:

Recall = TP / TP + FN

F1-score is defined as the harmonic average of recall and precision. The mathematical formula is defined as:

F1score=2*recall *precision/(recall+precision)

Where TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

The following results were obtained: -

- Accuracy

LEN OF REV	BER. NB	LINEAR SVC	LOG. REG	MULTI. NB	RANDOM FOREST
< 50	0.55	0.87	0.87	0.75	1.00
50–100	0.89	0.90	0.90	0.87	0.87
100 – 150	0.79	0.88	0.88	0.77	0.84
150 – 200	0.79	0.87	0.86	0.77	0.77
200 – 250	0.84	0.85	0.86	0.81	0.77
250 - 300	0.83	0.84	0.82	0.80	0.67

- Precision

LEN OF REV	BER. NB	LINEAR SVC	LOG. REG	MULTI NB	RANDOM FOREST
< 50	0.55	0.86	0.84	0.69	1.00
50–100	0.92	0.90	0.89	0.93	0.87
100 – 150	0.96	0.87	0.87	0.94	0.86
150 – 200	0.94	0.85	0.84	0.91	0.80
200 – 250	0.88	0.83	0.84	0.89	0.79
250 - 300	0.81	0.85	0.84	0.84	0.89

- Recall

LEN OF REV	BER. NB	LINEAR SVC	LOG. REG	MULTI NB	RANDOM FOREST
< 50	1.00	0.92	0.94	1.00	1.00
50–100	0.85	0.90	0.91	0.81	0.86
100 – 150	0.58	0.88	0.88	0.54	0.79
150 – 200	0.62	0.89	0.88	0.59	0.72
200 – 250	0.78	0.87	0.88	0.71	0.73
250 - 300	0.90	0.85	0.81	0.75	0.41

- F1 Score

LEN OF REV	BER. NB	LINEAR SVC	LOG. REG	MULTI NB	RANDOM FOREST
< 50	0.71	0.89	0.89	0.82	1.00
50–100	0.88	0.90	0.90	0.87	0.87
100 – 150	0.72	0.88	0.88	0.69	0.82
150 – 200	0.74	0.87	0.86	0.72	0.76
200 – 250	0.83	0.85	0.86	0.79	0.76
250 - 300	0.85	0.85	0.83	0.79	0.56

V. COMPARISON ANALYSIS OF FIVE ALGORITHMS:

After discussing the results, the authors decided to compare all the three algorithms based on the three parameters, i.e., precision, accuracy and F1- score. These three parameters are compared with the help of bar graph in order to display a perfect comparison. The comparisons are shown below as:

1. Accuracy:

The graph for accuracy is shown in the fig. 4.
We can observe that for reviews under 50 words, Random Forest algorithm performs the best with 100% accuracy however as the number of words increases its accuracy declines. On the contrary, Linear SVC and Logistic Regression consistently perform well and gives accuracy in the range 85-90% for all the varying lengths of the reviews. As for Bernoulli's Naïve Bayes and Multinomial Naïve Bayes the accuracy varies with the number of words in a group.

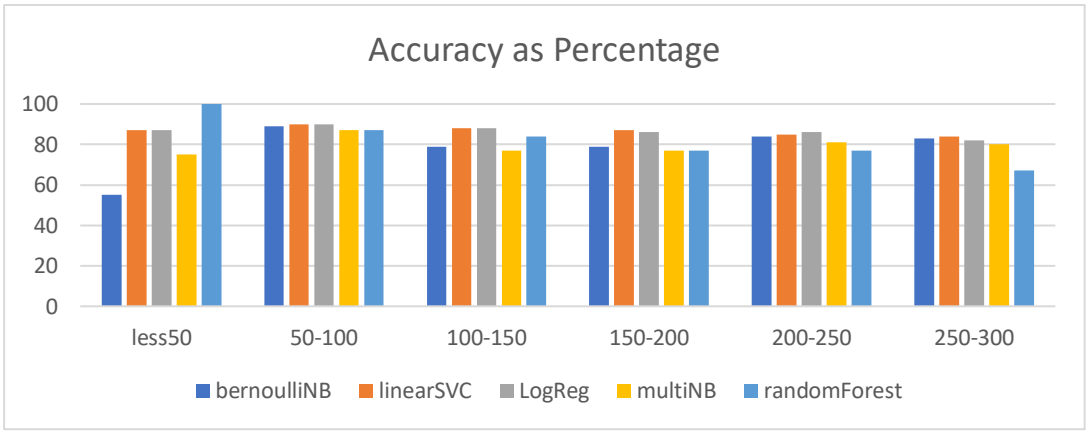


Fig.4 Categories versus Accuracy showing variationsof accuracy with respect to length of reviews in different algorithms

F1- score

The graph for accuracy is shown in the fig. 5.

F1 score is defined as the harmonic mean between precision and recall. It is used as a statistical measure to rate performance. In the above chart we can see that Linear SCV and Logistic Regression produce similar results for all the sets. Even though Random Forest performs extremely well in case of reviews with length less than 50 words, we can see a gradual decline as the length of the review increase.

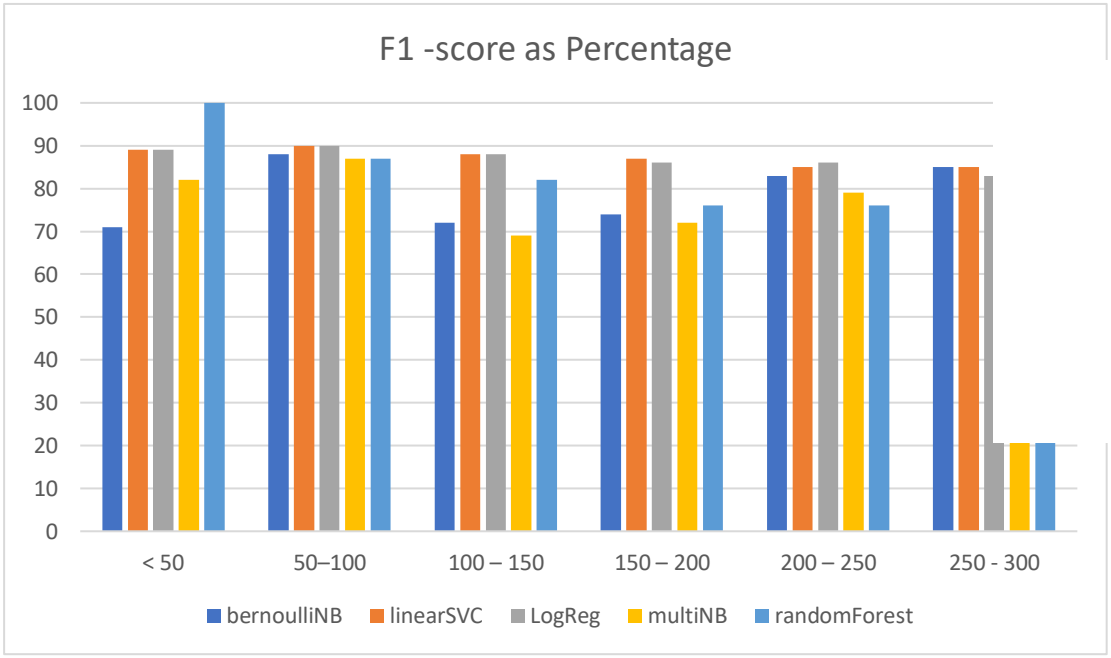
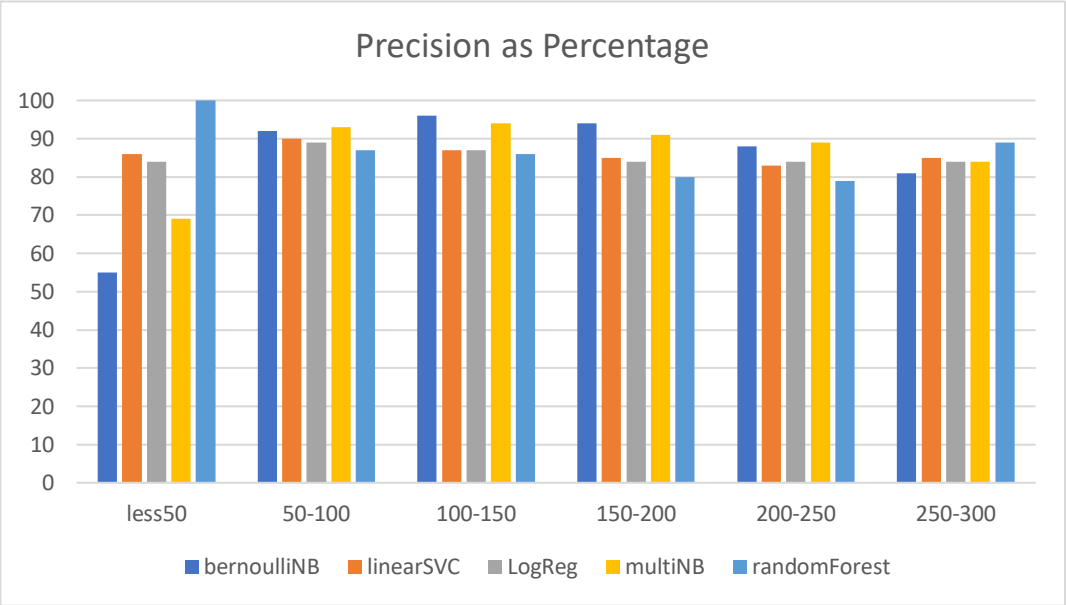


Fig.5 Categories versus F1 Score showing variations of accuracy with respect to length of reviews in different algorithms

2. Precision

The graph for accuracy is shown in the fig. 6.

We can observe that Bernoulli's Naïve Bayes has a very low precision for review with length less than 50 words. However, it performs the best overall for all the other groups. Logistic Regression and Linear SVC give similar precision for all the sets. As with other criterion, Random Forest performs the best for reviews having less than 50 words, but then its precision drops.



VI. Challenges and Future Scope:

The task of sentiment analysis is difficult due to the irregularity and complexity of language expressions. One of the biggest challenges of this task is brought about by hindered and negated expressions. These expressions contain a number of words that have a polarity which is opposite to the polarity of the expression itself. For example: 'Rohan performed well. His earlier two movies were overdramatic and slow and their plot were awful. However, the special effects made this movie spectacular.' In this statement, the number of negative words incorrectly implies that the statement is negative, when in reality it is actually positive and supportive. Although the model has been applied with excellent accuracy and precision rate, there are a few issues that need be taken into consideration for the work's future development.

When employing the SVM (support vector machine) algorithm, the data set does not give accuracy [11]. Additionally, the data set employed here is entirely text-based and statistical. Numerous applications in the actual world have shown significant success for random forest. We are currently dealing with the issue of learning from text data while there is a class imbalance (Wu et al. [12]). These algorithms' application space can include various data sets with features based on images and audios. POS (Part-Of-Speech) text recognition and picture recognition for the images data set are currently available technologies for these problems. These would give this research a wide range of applications if they were used. The usage of text classification applications can significantly advance the current trend toward automation.

VII. Conclusion:

This paper is constructing a IMDB Movie Review text classification model based on machine learning algorithms. This paper proposes the Linear SVC, Bernoulli Naive Bayes, Logistic Regression, Multinomial Naïve Bayes and Random Forest algorithms, which describes every aspect of model in detail by providing the evaluation metrics. When machine learning algorithms are implemented on a particular data set, the most important parameter that matters is the accuracy. Hence, the result shows that Random Forest classifier attains the highest accuracy of 100% for the reviews with less than 50 words. However, it fails to work efficiently if the number of words in the reviews increases. Thus, the algorithms that emerged as the most stable classifier which provided consistent high accuracy are Linear SVC classifier and Logistic Regression Classifier with accuracy between 80-90% through different review lengths.

VIII. References:

[1] Razzak, M.I., Naz, S., Zaib, A. (2018). Deep Learning for Medical Image Processing: Overview, Challenges and the Future. In: Dey, N., Ashour, A., Borra, S. (eds) Classification in BioApps. Lecture Notes in Computational Vision and Biomechanics, vol 26. Springer, Cham. https://doi.org/10.1007/978-3-319-65981-7_12

[2] L. Deng and X. Li, "Machine Learning Paradigms for Speech Recognition: An Overview," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 5, pp. 1060-1089, May 2013, https://doi.org/10.1109/TASL.2013.2244083

[3] Powers, J. G., Klemp, J. B., Skamarock, W. C., Davis, C. A., Dudhia, J., Gill, D. O., et al. (2017). The weather research and forecasting model: Overview, system efforts, and future directions. Bulletin of the American Meteorological Society, 98, 1717– 1737. https://doi.org/10.1175/BAMS-D-15-00308.1

[4] Prabhat A, Khullar V (2017) Sentiment classification on big data using Naïve bayes and logistic regression.In: Internationalconferenceon computer communication and informatics (ICCCI), pp 1–5

[5] Cherkassky, V., & Mulier, F. (1999). Vapnik-Chervonenkis (VC) learning theory and its applications. IEEE Transactions on Neural Networks, 10(5), 985-987. https://doi.org/10.1109/TNN.1999.788639

[6] Nadi A, Moradi H (2019) Increasing the views and reducing the depth in random forest. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2019.07.018

[7] Ranjitha KV (2018) Classification and optimization scheme for text data using machine learning Naive Bayes classifier. In: IEEE world symposium on communication engineering (WSCE), pp 33–36

[8] Yao H, Liu C, Zhang P, Wang L (2017) A feature selection method based on synonym merging in text classification system. EURASIP J WirelCommunNetw 2017:166. https://doi.org/10. 1186/s13638-017-0950-z

[9] Yuntao Z, Ling G, Yongcheng W, Yin Z (2003) An effective concept extraction method for improving text classification performance. Geo-Spatial Inf Sci 6(4):66–72

[10] Kumar R, Kaur J (2020) Random forest-based sarcastic tweet classification using multiple feature collection. In: Tanwar S, Tyagi S, Kumar N (eds) Multimedia big data computing for IoT applications. Intelligent systems reference library, vol 163. Springer, Singapore

[11] Liu Y, Loh HT, Tor SB (2005) Comparison of extreme learning machine with support vector machine for text classification. In: Ali M, Esposito F (eds) Innovations in applied artificial intelligence. IEA/AIE 2005. Lecture notes in computer science, vol 3533. Springer, Berlin, pp 390–399

[12] Wu Q, Ye Y, Zhang H, Ng MK, Ho S (2014) ForesTexter: an efficient random forest algorithm for imbalanced text Categorization. Knowl Based Syst 67:105–116