Twitter Sentiment Analysis Using Supervised Machine Learning

Nikhil Yadav, Omkar Kudale, Aditi Rao, Srishti Gupta, and Ajitkumar Shitole

Abstract Sentiment analysis aims to extract opinions, attitudes, as well as emotions from social media sites such as twitter. It has become a popular research area. The primary focus of the conventional way of sentiment analysis is on textual data. Twitter is the most renowned microblogging online networking site in which user posts updates related to different topics in the form of tweets. In this paper, a labeled dataset publicly available on Kaggle is used, and a comprehensive arrangement of pre-processing steps that make the tweets increasingly manageable to normal language handling strategies is structured. Since each example in the dataset is a pair of tweets and sentiment. So, supervised machine learning is used. In addition, sentiment analysis models based on naive Bayes, logistic regression, and support vector machine are proposed. The main intention is to break down sentiments all the more adequately. In twitter sentiment analysis, tweets are classified into positive sentiment and negative sentiment. This can be done using machine learning classifiers. Such classifiers will support a business, political parties, as well as analysts, etc., and so evaluate sentiments about them. By using training, data machine learning techniques correctly classify the tweets. So, this method doesn't require a database of words, and in this manner, machine learning strategies are better and faster to perform sentiment analysis.

N. Yadav (✉) · O. Kudale · A. Rao · S. Gupta · A. Shitole
International Institute of Information Technology, Pune, India
e-mail: nikhilyadav2698@gmail.com

O. Kudale
e-mail: omkark.py@gmail.com

A. Rao
e-mail: adirr0910@gmail.com

S. Gupta
e-mail: sg685641@gmail.com

A. Shitole
e-mail: ajitkumarsh1@gmail.com

J. Hemanth et al. (eds.), *Intelligent Data Communication Technologies and Internet of Things*, Lecture Notes on Data Engineering and Communications Technologies 57,
https://doi.org/10.1007/978-981-15-9509-7_51

Keywords Supervised machine learning · Sentiment analysis · Twitter · Data mining · Product evaluation · ROC · Classification · Naive Bayes · Logistic regression · Support vector machine · Linear SVC

1 Introduction

The Internet has been very useful in helping today's world express their perspectives globally. This is done through blog entries, on the web discussions forums, item audit sites, and so on. Individuals worldwide depend on this client, produced content extensively. For example, if someone wants to buy some product, then they first look up its reviews and comments before finalizing it [1]. But it is not humanly possible for a particular person to sit and look at every single review available. It would simply be a waste of time. Hence, to make this process easier, it can be automated. Machine Learning (ML) plays a significantly important part here. The process of Sentiment Analysis (SA) falling under ML helps the system understand the sentiment of a particular statement made. The system is built using several ML algorithms that can understand the nature of sentiment or a set of the same. In research, methods of ML have prevailed over knowledge- and dictionary-based methods to determine the polarity [2]. Polarity here is a semantic orientation that lies between two extremities, 0 and 1 or positive and negative. The paper proposes a system wherein data from twitter will be extracted on which SA will be performed. That data is saved in data frame. Then, some cleaning and pre-processing steps are performed on it so that, accurate information is utilized to fit the ML model which helps to predict labels for unknown cleaned and pre-processed data samples. Twitter is one of the popular sources containing a relatively huge amount of data. For performing sentiment analysis, certain supervised machine learning methods (algorithms) have been utilized to accomplish precise outcomes. Some of them are multinomial naive Bayes, linear support vector classifiers, and logistic regression classifiers. One is free to compose tweets in any form, without following any rules. This is what causes twitter more well known than other blogging locales. Due to this service, individuals tend to use abbreviations, make spelling mistakes, exaggerate reviews, use emoticons, etc., [3]. These formats usually make analysis a little difficult but still, there are methods such as feature extraction and mapping emoticons to their actual meanings that can be utilized to investigate the tweets. Movie and item audit easily accessible nowadays or thoughts on religious and political issues, so they become fundamental wellsprings of client slant and sentiment. This paper majorly focuses on data that are related to product reviews for product evaluation. It is restricted mainly to the data of vendors, manufacturers, entrepreneurs, and others of a similar domain. Messages can change from general opinion to individual idea [4].

2 Related Work

Sentiment analysis is the careful examination of how feelings and points of view can be identified with one's feeling and mentality appears in regular language regard to an occasion. The principle motivation behind choosing twitter's profile information is that subjective data can get from this platform [5]. Ongoing occasions show that sentiment analysis has reached incredible accomplishment which can outperform the positive versus negative and manage the entire field of behavior and feelings for various networks and themes. In the field of sentiment analysis utilizing various techniques, great measure of exploration has been done for the expectation of social sentiments. Pang and Lee (2002) proposed the framework, where an assessment can be positive or negative was discovered by the proportion of positive words to total words. Later in 2008, the creator built up a methodology in which tweet results can be chosen by term in the tweet [6]. Another study on twitter sentiment analysis was done by Go et al. [7] who stated the issue as a two-class classification, meaning to characterize tweets into positive and negative classes. M. Trupthi, S.Pabboju, and G.Narasimha proposed a system that mainly makes use of Hadoop. The data is extracted using SNS services which are done using twitter's streaming API. The tweets are loaded into Hadoop and are pre-processed using map-reduce functions. They have made use of uni-word naive Bayes classification [8]. The paper [9] analyzes the utilization of SA in business applications. Besides, this paper exhibits the text analysis process in auditing the popular assessment of clients toward a specific brand and presents hidden information that can be utilized for decision making after the text analysis is performed. In paper [10], the sentiment analysis has been done in four phases. Collecting real-time tweets up to a given limit, tokenizing every tweet as part of pre-processing, comparing them with an available bag of words, and classifying the tweets as positive or negative.

The proposed system is domain specific. A user interactive GUI will be available for the users to type in the keywords related only to the commercial products. Not many existing systems have been made so specific. Also, the system aims to compare various ML algorithms and choose the one which will produce results with the highest accuracy. Making the system domain specific reduces processing time as tweets regarding specific products are only searched based on the keywords typed.

3 Proposed System

The system intends to carry out sentiment analysis over tweets gathered from the twitter dataset. Various algorithms have been utilized and tested against the available dataset, and the most appropriate algorithm has been chosen. Figure 1 gives the idea about how the sentiment analysis will be carried out. Once the dataset has been cleaned and divided (isolated) into preparing (training) and testing datasets, it will be pre-processed using the techniques mentioned below. Features will be extracted

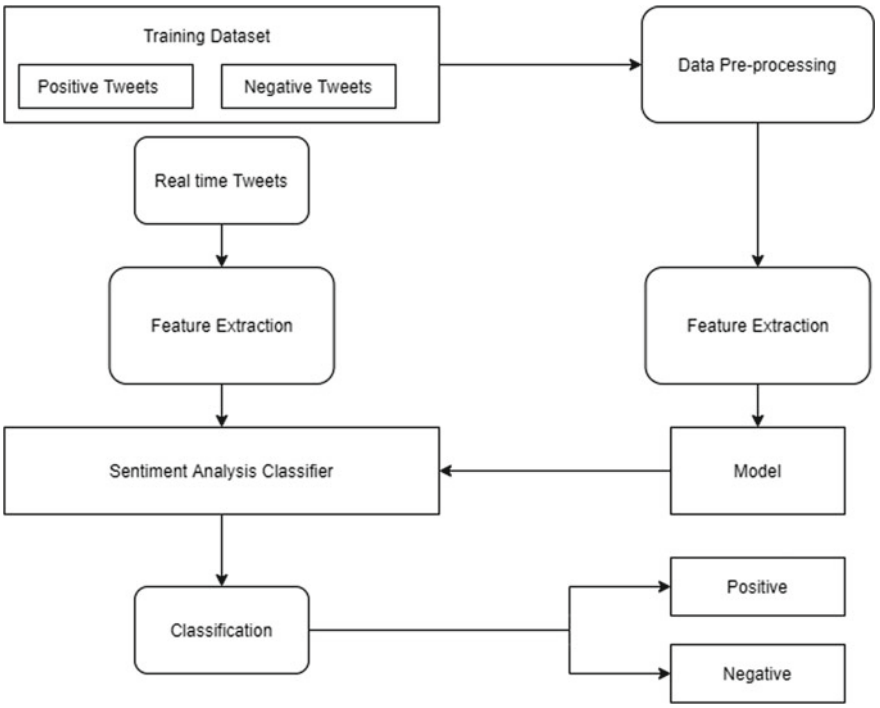

Fig. 1 Outline of proposed system

to reduce the dimension of the dataset. The next stage is to create a model that will be given to the classifier to classify the tweets into positive and negative tweets. Again real-time tweets will be given to the classifier for testing the real-time data. The proposed system does not engage in performing sentiment analysis on every tweet belonging to every other domain. The system is strictly domain restricted, where the sentiment analysis is performed to classify the tweets related to products in the market into a negative or positive category. The end-user will be provided with an interactive GUI wherein he/she can type the keywords or sentences related to a particular product. All tweets which are identified with that product will be available to the user. The user will be able to see the number of positive and negative statements made by others. This will help them in revising their production and work strategies accordingly which will be useful in improving their businesses.

Below are steps involved in handling large incoming data:

1. Data cleaning:
   i. Use of various data tools that can help in cleaning the dataset.
   ii. Use of several AI tools that help in identifying duplicates in large corpora of data and eliminate it.
   iii. For correcting the corrupted data, the source of errors should be tracked and monitored constantly.

iv. Validate once, when the existing data is cleaned.
2. Data pre-processing:
i. Assessing data quality.
ii. Identification of inconsistent values to know what the data type of the features should be.
iii. Aggregate the features to give better performance.

Sections 3.1 and 3.2 will give detailed ideas to handle large incoming data.

3.1 Data Cleaning

The data collected was not in the correct format. Information cleaning is the way toward guaranteeing that information is right, predictable, and usable. Usually, datasets need to cleanse because they consist of a lot of noisy or unwanted data called outliers too. The existence of such outliers may lead to inappropriate results. Data cleaning ensures the removal and improvisation of such data and results in a much more reliable and stable dataset.

Data cleaning can be done in given ways:

- **Monitors errors.** The entry point or source of errors should be tracked and monitored constantly. This will help in correcting the corrupted data.
- **Process standardization**. The point of entry should be standardized. By standardizing the data process, the risk of duplication reduces.
- **Accuracy validation**. Data should be validated once the existing database is cleaned. Studying and using various data tools that can help in cleaning the datasets is very important.
- **Avoid data duplication**. The identification of duplicate data is a very mandatory process. Several AI tools help in identifying duplicates in large corpora of data.

The above-mentioned steps are a few of the many ways to clean datasets. Making use of these methods will end up giving good, usable, and reliable datasets.

3.2 Data Pre-processing

The next step after data cleaning is data pre-processing. It is a major step in machine learning. It is the process in which data gets transformed or encoded to a state which is understandable to the machine. In simple words, the features of the dataset can be easily interpreted by the algorithms. The feature is a measurable property of an entity being observed. For example, height, age, gender can be considered as features of a person. A twitter stream will extract every related tweet from twitter, which will be in an unstructured structure. These unstructured tweets need to experience pre-handling before applying any classifier over it. The tweets will be pre-handled with

tokenization and cleaning. Initially, all the HTML contents are expelled from the tweets by giving a URL structure. Further, cleaning happens by expelling non-letters or images using python regular expressions. All the tweets must be in the same case to processes; thus, it will change over to bring the lower case, and each word is split based on space. Following this, gather all stop words and structure it as a solitary set and evacuate it. At last, return a string of important words [11]. Thus, a pre-processing step is performed for filtering out the slang words and misspellings before extracting the features[12]. Following steps can help with data pre-processing:

- **Data quality assessment**. Since the data is collected from multiple sources, it will be unrealistic to consider it to be perfect. Assessing the data quality must be the first step while pre-processing it.
- **Inconsistent values**. Data can be inconsistent at times. Like the "address" field can contain a phone number. Hence, the assessment should be done properly like to know what the data type of the features should be.
- **Feature aggregation**. As the name says, features are aggregated to give better performance. The behavior of aggregated features is much better when compared to individual data entities.
- **Feature sampling**. It is a way of selecting a subset of the original (first) dataset. The central matter of sampling is that the subset should have nearly the same properties as the original dataset.

Coming to the proposed system, the pre-processing done will be as follows:

- Converting tweets to lowercases.
- Supplant at least two dots with spaces.
- Replace extra spaces with a single one.
- Get rid of spaces and quotes at the end of the tweets.

Two types of features are extracted from the dataset, namely unigrams and bigrams. A frequency distribution is created for the extracted features. Later, the top N unigrams and bigrams are chosen to carry out the analysis. Also, tweets contain special features like URLs, user names, emoticons, etc. Retweets are also a feature of tweets. These features are not required while performing sentiment analysis. Hence, these features are replaced with common keywords or markers like “URL,” “USER_MENTION,” “EMO,” respectively. Again, removal of stop words and lemmatization are necessary steps to be done.

Stop words. Stop words will be words that don’t have any criticalness in search inquiries. For example, “I like to write.” After removing stop words becomes, “like write.” “I” and “to” are termed as stop words.

Stemming. It is an element procedure of delivering morphological variations of a base word. The words like “chocolatey,” “chocolates” are converted to their root word “chocolate.”

Lemmatization. Lemmatization decreases the inflected words appropriately guaranteeing that the root word has a place with the language.

3.3 Classifiers to Be Used

Naive Bayes. The naive Bayes is a supervised machine learning algorithm that returns probability values as the output. Naive Bayes classifier is very useful in solving high-dimensional problems. It assumes the probabilities of the different events that are completely independent. Naive Bayes [13] is a straightforward model, where class C is assigned a tweet t such that:

$$C = \arg\max P(c|t) \tag{1}$$

$$P(c|t) \infty P(c) \prod^{n} P(fi|c) \tag{2}$$

The probability of event A happening can be found, by giving the occurrence of event B. Naive Bayes algorithm can be used to tackle large scale classification problems.

Logistic regression. Logistic regression predicts a binary outcome, i.e., (Y/N) or (1/0) or (True/False). It also works as a special case of linear regression. It produces an S-shaped curve better known as a sigmoid. It takes real values between 0 and 1.

The model of logistic regression is given by:

$$Output: 0 \text{or} 1 \tag{3}$$

$$\text{Hypothesis: } Z = WX + B \tag{4}$$

$$h_\theta(x) = \text{sigmoid}(Z) \tag{5}$$

Basically, logistic regression has a binary target variable. There can be categories of target variables that can be predicted by it. The logistic classifier uses a cross-validation estimator.

Support vector machine. It is a non-probabilistic model that utilizes a portrayal of text models as focuses in a multidimensional space. These examples are mapped with the goal that the instances of the diverse categories (sentiments) have a place with particular areas of that space. Later, the new messages are mapped onto that equivalent space and are predicted to have a place with a classification dependent on which category they fall into. In the SVM algorithm, the fundamental goal is to boost the edge between information points and the hyperplane. The loss function that helps with this is called a hinge loss. The equation of the hyperplane is given as:

$$w.x - b = 0 \tag{6}$$

$$c(x, y, f(x)) = \begin{cases} 0, if y * f(x) \geq 1 \\ 1 - y * f(x), \text{else} \end{cases} \tag{7}$$

The cost is 0 if the predicted, and the actual value is of a similar sign. On the off chance that they are not, at that point, figure the loss esteem.

For performing sentiment analysis, logistic regression is considered over naive Bayes because naive Bayes assumes all the features used in model building to be conditionally independent whereas logistic regression splits feature space linearly and typically works reasonably well even when some of the variables are correlated, and on the other hand, logistic regression and SVM with a linear kernel have similar performance but depending on the features, one may be more efficient than the other.

3.4 Plotting Results

The plotting of the result will be done using graphs, and the comparison of algorithms is done using a ROC curve [14]. It is a plot of true positive rate and false positive rate. Figure 2 shows the ROC curve for a MultinominalNB Model. The Area Under the Receiver Operating Characteristics curve (AUROC) for the MultinominalNB Model is 0.89.

Figure 3 shows the ROC curve for a logistic regression model. The Area Under the Receiver Operating Characteristics curve (AUROC) curve for the logistic regression model is 0.90.

Figure 4 shows the ROC curve for a linear SVC model. The Area Under the Receiver Operating Characteristics curve (AUROC) for the linear SVC model is 0.83.

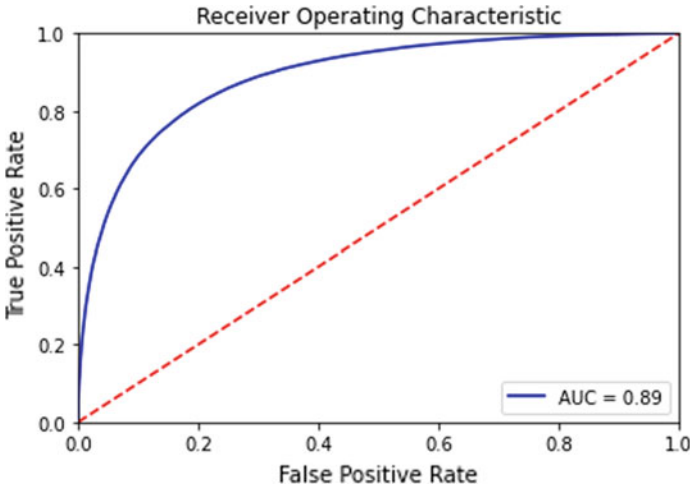

Fig. 2 ROC curve for multinomial naive Bayes classifier

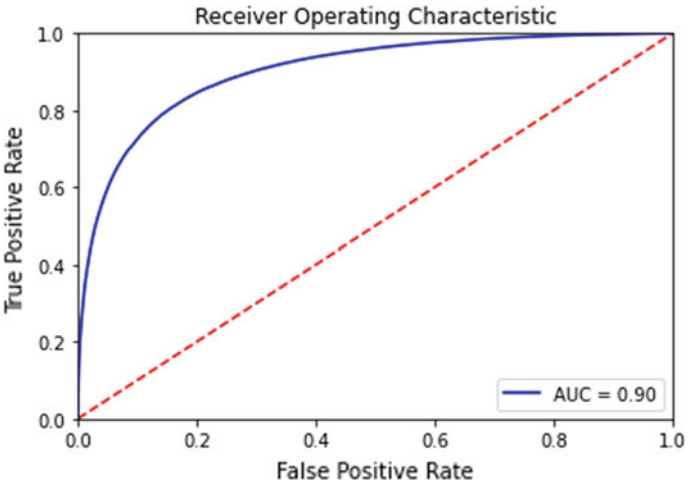

Fig. 3 ROC curve for logistic regression classifier

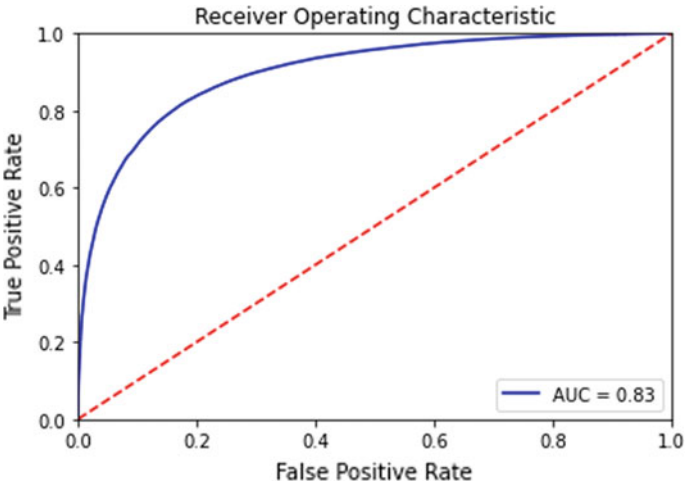

Fig. 4 ROC curve for linear SVC

4 Results

The dataset used for the training model is the Sentiment140 dataset. It is a balanced dataset with 1.6 million tweets among which 8 lakh tweets belong to the positive class, and the remaining 8 lakh tweets belong to negative class. The splitting is done using the train_test_split method with a test_size of 0.20. 12 lakh tweets are used for training the model, and the remaining 4 lakh tweets are used for testing the model.

Table 1 Classification report of logistic regression

	Precision	Recall	F1 Score	Support
0 (Negative label)	0.81745	0.83662	0.82692	160,156
1 (Positive label)	0.83236	0.81280	0.82246	159,844
Accuracy			0.82472	320,000
Macro avg	0.82490	0.82471	0.82469	320,000
Weighted avg	0.82490	0.82472	0.82470	320,000

Table 2 Classification report of multinomial naive Bayes

	Precision	Recall	F1 Score	Support
0 (Negative label)	0.78090	0.85148	0.81466	160,156
1(Positive label)	0.83637	0.76064	0.79671	159,844
Accuracy			0.80610	320,000
Macro avg	0.80864	0.80606	0.80569	320,000
Weighted avg	0.80861	0.80910	0.80569	320,000

4.1 Logistic Regression

Table 1 gives the classification report of the logistic regression model. The accuracy of the model is 82.47. It also shows the precision and recall of the model. Precision is the positive predictive value, and recall is the sensitivity of the model [15].

4.2 Multinomial Naive Bayes

Table 2 gives the classification report of multinomial naive Bayes model. The accuracy of the model is 80.61.

4.3 Linear Support Vector Machine

Table 3 shows the classification report of the linear SVC model. The accuracy of the model is 83.71.

ROC curves are appropriate when the observations are balanced between each class and since the dataset used for training and testing is a balanced dataset ROC curves which are considered for measuring the performance of the model.

AUROC is a superior measure of classifier performance than accuracy because it does not bias on size of test or evaluation data whereas accuracy is always biased on size of test data, and also AUROC is the best summary of the performance of a

Table 3 Classification report of linear support vector machine

	Precision	Recall	F1 Score	Support
0 (Negative label)	0.85970	0.90315	0.88889	160,011
1 (Positive label)	0.78447	0.70515	0.74270	79,989
Accuracy			0.83716	240,000
Macro avg	0.82288	0.88415	0.81179	240,000
Weighted avg	0.83462	0.83716	0.83483	240,000

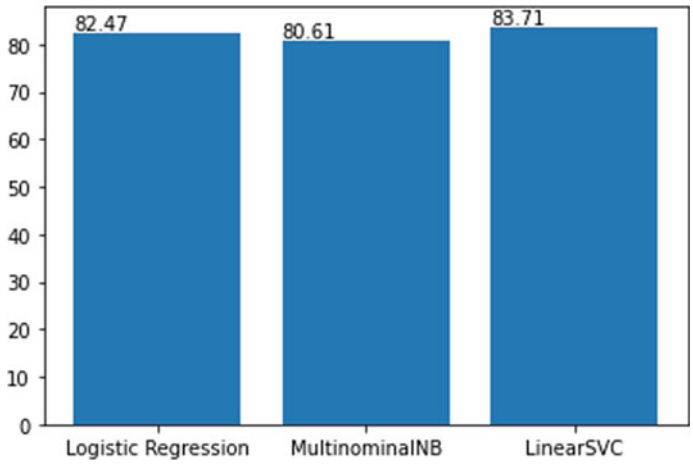

Fig. 5 Comparison of accuracies of classifiers

classifier as it incorporates different aspects of the performance into a single number. Figure 5 shows a comparison between the three algorithms used for sentiment analysis; comparatively, linear SVC gives the highest accuracy of 83.71 but its AUROC is less than logistic regression having an accuracy of 82.47, and hence, logistic regression is considered for classification purpose.

5 Conclusion and Future Work

The work in this paper is done to classify a relatively huge corpus of twitter data into two groups of sentiments, positive and negative, respectively. Higher accuracy is achieved by using sentiment features instead of conventional text classification. This feature can be used by various establishments, business organizations, entrepreneurship, etc., to evaluate their products and get a deeper insight into what people say about their products and services. Future work includes working not only in the English language but in other regional languages too. Also, it will include analysis

of complex emotions like sarcasm and generate a hybrid classifier to get the best accuracy.

References

1. Neethu MS, Rajasree R (2013) Sentiment analysis in twitter using machine learning techniques. In: 2013 Fourth international conference on computing, communications and networking technologies (ICCCNT), Tiruchengode, pp 1–5
2. Kumar A, Sebastian TM, Sentiment analysis on twitter. Department of Computer Engineering, Delhi Technological University Delhi, India
3. Joshi R, Tekchandani R (2016) Comparative analysis of Twitter data using supervised classifiers. In: 2016 International conference on inventive computation technologies (ICICT). ISBN: 978-1-5090-1285-5
4. Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau R, Sentiment analysis of twitter. Passonneau Department of Computer Science Columbia University New York, NY 10027 USA
5. Hasan A, Moin S, Karim A, Shamshirband S, Machine learning-based sentiment analysis for twitter accounts. Department of Computer Science, Air University, Multan Campus, Multan 60000
6. Pang B, Lee L (2008) Opinion mining and sentiment analysis
7. Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford 1, no 12
8. Trupthi M, Pabboju S, Narasimha G (2017) Sentiment analysis on twitter using streaming API. In: 2017 IEEE 7th international advance computing conference (IACC), Hyderabad, pp 915–919
9. Halibas AS, Shaffi AS, Mohamed MAKV (2018) Application of text classification and clustering of twitter data for business analytics. In: 2018 Majan international conference (MIC), ISBN: 978-1-5386-3761-6
10. Neethu MS, Rajasree R (2013) Sentiment analysis in twitter using machine learning techniques. In: 2013 Fourth international conference on computing, communications and networking technologies (ICCCNT), Tiruchengode, pp 1–5
11. Shamantha RB, Shetty SM, Rai P (2019) Sentiment analysis using machine learning classifiers: evaluation of performance. In: 2019 IEEE 4th international conference on computer and communication systems (ICCCS), Singapore, pp 21–25
12. Tyagi P, Tripathi RC (2019) A review towards the sentiment analysis techniques for the analysis of twitter data. In: Proceedings of 2nd international conference on advanced computing and software engineering (ICACSE)
13. Yadav N, Kudale O, Gupta S, Rao A, Shitole A (2020) Twitter sentiment analysis using machine learning for product evaluation. In: 5th International conference on inventive computation technologies (ICICT-2020)
14. Shitole A, Devare M (2019) TPR, PPV and ROC based performance measurement and optimization of human face recognition of IoT enabled physical location monitoring. Int J Recent Technol Eng 8(2):3582–3590. ISSN: 2277-3878
15. Shitole A, Devare M (2018) Optimization of person prediction using sensor data analysis of IoT enabled physical location monitoring. J Adv Res Dyn Control Syst 10(9):2800–2812. ISSN: 1943-023X