Multimedia Tools and Applications (2022) 81:42261–42283
https://doi.org/10.1007/s11042-022-13492-w

1220: VISUAL AND SENSORY DATA PROCESSING FOR REAL TIME INTELLIGENT SURVEILLANCE SYSTEM

Sentiment analysis of COVID-19 social media data through machine learning

Dharmendra Dangi[1] · Dheeraj K. Dixit[1] · Amit Bhagat[1]

Received: 24 February 2021 / Revised: 15 October 2021 / Accepted: 13 July 2022 /
Published online: 25 July 2022

Abstract

Pandemics are a severe threat to lives in the universe and our universe encounters several pandemics till now. COVID-19 is one of them, which is a viral infectious disease that increased morbidity and mortality worldwide. This has a negative impact on countries' economies, as well as social and political concerns throughout the world. The growths of social media have witnessed much pandemic-related news and are shared by many groups of people. This social media news was also helpful to analyze the effects of the pandemic clearly. Twitter is one of the social media networks where people shared COVID-19 related news in a wider range. Meanwhile, several approaches have been proposed to analyze the COVID-19 related sentimental analysis. To enhance the accuracy of sentimental analysis, we have proposed a novel approach known as Sentimental Analysis of Twitter social media Data (SATD). Our proposed method is based on five different machine learning models such as Logistic Regression, Random Forest Classifier, Multinomial NB Classifier, Support Vector Machine, and Decision Tree Classifier. These five classifiers possess various advantages and hence the proposed approach effectively classifies the tweets from the Twint. Experimental analyses are made and these classifier models are used to calculate different values such as precision, recall, f1-score, and support. Moreover, the results are also represented as a confusion matrix, accuracy, precision, and receiver operating characteristic (ROC) graphs. From the experimental and discussion section, it is obtained that the accuracy of our proposed classifier model is high.

Keywords Logistic regression · Random forest · Multinomial Naïve Bayes · Support vector machine · Decision tree · COVID-19

✉ Dharmendra Dangi
dangi28dharmendra06@gmail.com

[1] Department of Mathematics, Bioinformatics and Computer Applications, Maulana Azad National Institute of Technology, Bhopal, India

1 Introduction

Coronavirus disease 2019 (COVID-19) was discovered in Wuhan, China after the virus had spread globally. It was announced by the World Health Organisation to be a pandemic. This epidemic is now affecting a large number of individuals all across the world. At present, COVID-19 is a serious threat to human life all over the world, where several individuals developed symptoms such as pneumonia [46]. It has a wide variety of effects on the human body, including extreme respiratory syndrome and multi-organ failure, which will potentially lead to death within a short period of time [32]. When the globe has been fighting for COVID-19 in recent months and most people have been imprisoned, Twitter has become more important than ever. Even in the past, people have been using Twitter to communicate, express, and spread information relevant to the disaster, whether it be cyclones, ebola, flooding, or Zika [4, 8, 26, 42, 48]. Twitter has been one of the platforms for millions to express their emotions regarding different issues.

Social networking has been a crucial public platform for data acquisition and social learning to manage uncertainties and risks during a national epidemic. X. Gui et al. [12] investigated Public issues about the Zika virus epidemic, and personal risk management processes and travel-related judgment during the epidemic have been identified. People use Twitter to express their opinions and to spread information. The rapid exchange of consumer opinions on social media has allowed researchers to recognize sentiments about almost everything, including thoughts about goods, films, politics, new technologies, and natural disasters [19].

Nowadays, the worldwide economic crisis has done due to this Pandemic. Many sectors, industries, education systems, and all other sections are suffering due to COVID-19. Nowadays, health workers, doctors, and peoples share lots of knowledge on social networks like Twitter, social media, Facebook, etc. Moreover, social networks are the most vital medium to explore the intrinsic details about COVID-19. Here we are analyzing social media COVID-19 data through social network platforms. Several experiments are conducted using machine learning methods such as coronary heart disease to predict multiple diseases [5, 10, 11, 13–16, 20, 23, 30, 31, 34, 35, 43–45, 49].

So Machine learning (ML) algorithms have proved to get a sentiment analysis of COVID-19 types of pandemics issues. In this work, a research experiment has been conducted to analyze the sentiments of Citizens of India towards COVID-19. It is to classify what emotions people from various regions of this country have been expressing. We have proposed a novel Sentimental Analysis of Twitter social media Data (SATD) approach which utilizes five classifiers as Logistic Regression, Random Forest Classifier, Multinomial NB Classifier, Support Vector Machine, and Decision Tree Classifier. The proposed approach is used to classify the tweets into three class indexes. This method effectively classifies the tweets with high accuracy, precision, recall, and F1-score. The main contributions of this work are presented as follows:

- This paper presents five dinstinct machine learning models namely Random Forest Classifier, Multinomial NB Classifier, Logistic Regression, Support Vector Machine, and Decision Tree to accurately identify the sentiments present in COVID-19 related tweets.
- The main aim of this work is to test the effectiveness of these classifiers for processing a large number of features.

- The usage of five machine learning classifiers offers high performance on the sentimental analysis task and these models are not prone to overfitting and are biased by outliers.
- For the evaluations and comparisons, the twint dataset has been utilized with different sentiment classes (Positive, negative, and neutral).

The rest of the work is organized accordingly; in section 2, the relevant works of the proposed work are reviewed. Section 3 describes the different machine learning approaches, the data collection & labeling and the proposed methodology are explained in section 3. The experimental analysis is depicted in section 4. Finally, the work is summarized in section 5.

2 Review of related works

Researchers had contributed a lot in the area of COVID-19 analysis. However, all current research in this area has been designed for statistical and machine learning models to predict useful information regarding this disease. Ndaïrou et al. [29] proposed a statistical and mathematical model for COVID-19 disease. This model explores the propagation of COVID-19 disease. They focused on transmissibility among individuals or in a group. Based on the study's threshold value, for accomplishing these tasks, they used the simulation model COVID-19 dataset for the outbreak that occurred in Wuhan, China. Zhang et al. [51] Analysis of COVID-19 daily news case details for six Western countries such as Canada, France, Germany, Italy, the United Kingdom, and the United States. To evaluate such information, they have a Poisson model. This research allowed us to make a statistical forecast of different parameters, such as attack rate.

Rustam et al. [36] used forecasting mechanisms by using machine learning. In the potential course of action, they have reinforced the decision-making mechanism. Linear regression, support vector machine, least absolute shrinkage, selection operator, and exponential smoothing are the four standard prediction models for predicting the data model. These models are being used to evaluate the COVID-19 dataset. Through the study of COVID-19 datasets, these perditions provide various findings such as number of fatalities, amount of recoveries, number of newly infected cases.

Sear et al. [40] use machine learning techniques with various parameters like health guidance components, vaccinations, etc. They found the pro-vaccination community is more focused on debate compare to the anti-vaccination community. They developed a model for analyzing these kinds of data. They have also demonstrated that their model is scalable, thereby solving the urgent issue facing social media sites of evaluating large amounts of misinformation and disinformation regarding online wellbeing. Li, Sijia, et al. [24] have been observed an increase in negative feelings (anxiety, sadness, and resentment) and susceptibility to social threats, and a reduction in positive emotions and overall happiness after declaration COVID-19 is a Pandemic. The author also discusses the effect of COVID-19 on people's mental health, supports decision-makers in designing actionable strategies, and assists healthcare professionals, such as social workers, therapists, and psychologists, in delivering timely care to impacted communities. So they found that COVID-19 has profoundly led to a vast number of psychological effects. Wang et al. [25] carried out a process of cognitive analysis of COVID-19 data obtained from Sina Weibo, a social news platform in China using Support Vector Machine, Naive Bayes, and Random Forest Classifiers.

3 Proposed methodology

3.1 Machine learning techniques

There are various applications of ML rather than modeling COVID-19 types of pandemics. Moreover, various other real-world applications are driver-free vehicles, business applications, natural language processing (NLP), Medical, Banking, E-commerce, etc. ML is widely used for forecasting or predictions such as disease forecasting, weather forecasting, e-commerce, and disease prediction. Doing this can make different types of machine learning models like regression, neural network models, etc. Machine learning algorithms are very useful in various areas such as E-commerce, Social Network Analysis Medial, etc. [6, 27, 28]. There are three types of ML techniques such as supervised, unsupervised, and semi-supervised learning.

In supervised learning, datasets are broken down into two parts, i.e., training datasets and test datasets. In the sentiment analysis of Twitter social media data, preprocess that Twitter social media data and find the class labels. After that, analyze the Twitter social media data by using the ML algorithm. In unsupervised learning, labeled training datasets are not used instead of this direct input of datasets that are provided to the system [38, 41, 47]. Once the training phase is completed, then test on test data sets can be performed.

In Semi-Supervised Learning uses a mixer of labeled and unlabeled data. This kind of technology system requires a minimum amount of labeled data. So with less effort, more efficient results can be produced using semi-supervised learning [1, 33]. Various machine learning algorithms can be used in the Twitter social media data analysis, such as well-liked algorithms are Decision Tree, Support Vector Machine (SVM), Naive Bayes Classifier, Clustering, etc. [18, 37, 39]. Here we have used 5 machine learning classifiers for predicting and analyzing the data. Such classifiers are Logistic Regression, Random Forest Classifier, Multinomial NB Classifier, Support Vector Machine, and Decision Tree [21, 22]. Figure 1, proposed an ML model for Twitter social media data based upon COVID-19.

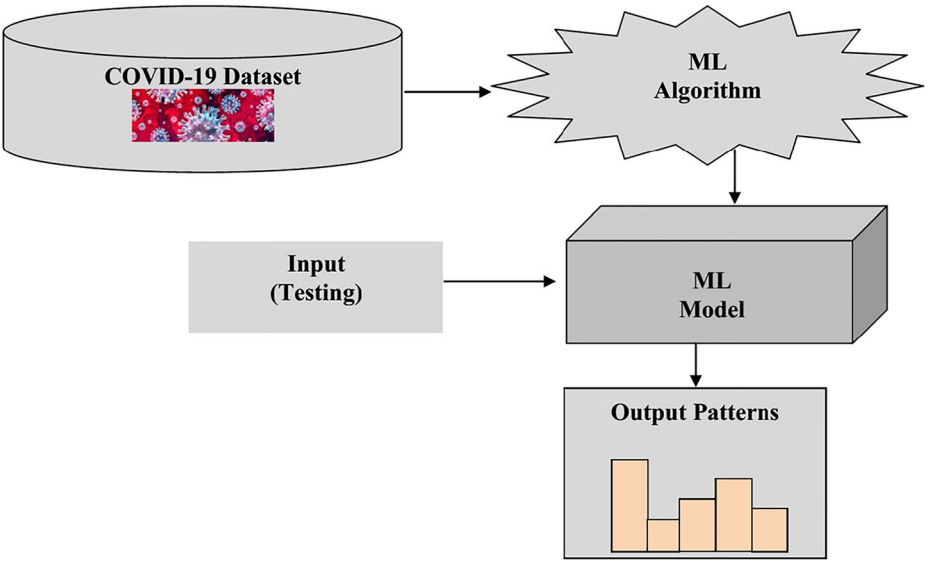


Fig. 1 ML model for COVID-19 data analysis

3.1.1 Random forest classifier

This is a supervised learning algorithm, and this classifier uses both classification and regression. Generally, it is used for classification. Moreover, it creates a decision tree of data and predicts the best solution using voting. It is a collection of functions that is as good as a single decision tree because it reduces the over-fitting by averaging the result [9]. This is actually a multiple-trained decision tree classifier and uses N number of decision tree classifiers. The input tweets are transmitted via each decision tree and thereby obtain the class indexes. The results obtained from the trees are sent to the bagging technique and thus the tweets are classified.

3.1.2 Multinomial Naïve Bayes (MNB)

It utilizes the relative occurrence of a word in documents belonging to the class, the conditional likelihood of a given class. The classification tweet based on the classes is performed by using the MNB and can be given as,

$$P(Q|S) = \frac{P(S|Q)P(Q)}{P(S)} \tag{1}$$

Here, Q denotes the COVID-19 related tweets progression, and S is the number of COVID-19 related tweets. The features are independent to each other and thus the name Naïve.

3.1.3 Logistic regression

It is a classifier which is being used in the ML algorithm for classification problem. It uses the statistics and probability theory for analysis [50]. The probability of P of dichotomous result can be expressed in the form,

$$\log it(P) = \ln\left[\frac{P}{1-P}\right] = \alpha_0 + \alpha_1 y_1 + \ldots. + \alpha_n y_n = \alpha_0 + \sum_{i=1}^{n} \alpha_i y_i \tag{2}$$

The coefficients $\alpha_1, \alpha_2, \ldots, \alpha_n$ incorporated with the explanatory variable $y_1, y_2, \ldots, y_n$. α_0 is the intercept.

3.1.4 Support vector machine

It is a classifier that refers to an ML algorithm that is supervised. It uses classification and regression data to be analyzed. Moreover, it can be used to attain classification of tweets based on the index more accurately and consumes less power than the other approaches. SVM focused to find the hyperplane in an N-D space and thereby classifies the tweets. The hyperplanes are used as a decision boundary and can be used to classify the tweets, and N is the number of features. The effectiveness of the classification accuracy relies on the margin of the hyperplane and if it is maximum then the tweets are classified distinctively.

3.1.5 Decision tree

This is a supervised machine learning classifier in which the tweets are segmented continuously based on the available parameters. This classifier is used for prediction and classification. It uses the tree data structure for analyzing the data. Each internal node in the decision tree

represents a test on an attribute, the branch displays the test result, and the class mark is shown by a leaf or terminal [7]. The classified classes Neutral Sentiment- Index 0, Positive Sentiment – Index 1, and Negative Sentiment- Index 2 are available in the leaf node.

There are various tools used for data preprocessing and machine learning, such as Twint- it is a Twitter social media intelligence tool. It is an advanced social media scraping and OSINT Twitter tool written in Python that does not use the API of Twitter social media, allowing you to scrape the followers of a person, subscribe, Comment, and more while ignoring most API limitations. Python contains a built-in module called Python Data Analysis Library (Pandas) that aids in the smooth flow of operations. Pandas is a popular Python library for data processing activities including cleaning, modifying, and analyzing data. It contains classes that let you read, process, and write CSV data files. Although there are several data cleaning methods available, the Pandas library provides a highly fast and effective way to manage and analyze data. It does so by providing us with Series and DataFrames, which allow us to not only properly represent data but also to manipulate it in many ways.

NLTK - It is the Natural Language Toolkit, a compilation of open-source software modules, instructions, and problem sets, offering computational linguistics course content ready for use.

Textblob- TextBlob is a python library for processing textual data.

Scikit-Learn- It is an open-source library for machine learning that facilitates supervised and unsupervised learning. It also offers numerous model-fitting methods, data preprocessing, model choice and assessment, and several other utilities.

Matplotlib- It is a vast library in which Python can construct static, animated, and immersive visualizations.

Seaborn- Seaborn is a library of matplotlib-based Python data visualization. It offers a high-level gui for convincing and insightful statistical graphics.

3.2 Data collection and labeling

There are several methods of gathering data for social media sentiment analysis. Social media platforms, such as Twitter, provide user tweets for the public that are open for research. You will explicitly import tweets from the Twitter website. To perform multiclass classification, we have observed tweets related to the covid-19 pandemic. Twints are used to collecting tweets from Twitter, and these tweets belong to Indian Twitter handle using different keywords like COVID-19, Corona, Pandemic, and many more. Twin tools work as Twitter API. It is required a consumer key, Consumer secret, access token, and access token secret key. Here we have to use the Python Language interface to access tweets using these keys [3].

These Two Twitter datasets are taken from social media sites. In which the first dataset holds the tweet during the lockdown. In contrast, the second dataset contests the tweets after the completion of the post lockdown. With these two different datasets, we will try to understand the sentiments of the public. Raw data was retrieved using feature analysis tweets using various queries such as covid-19, COVID, corona, coronavirus, pandemic, and tweet intersection containing Wuhan and virus. Now a day researchers are interested in analyzing the opinion of COVID-19 tweets [2]. Before fitting the machine learning model, we have to preprocess the data, which basically has some basic steps.

- ***Tokenization:*** It can be accomplished by splitting documents (crawled reviews) into a list of tokens such as words, numbers, special characters, etc., and make the document ready to be used for further processing.

- ***Normalization:*** This process converts all the word tokens of a document into either lower case or upper case because most of the reviews consist of both cases i.e. lower-case and uppercase characters. The purpose of shifting all tokens into a single format can easily be used for prediction.
- ***Stop Word Removal:*** In this step, we have removed all the commonly occurring words from each of the tweets. For removing stop words, a predefined list of stop words has been used. Each tweet is compared with the available stop word list, and matching words are removed from that tweet. These words do not contribute to the performance of the model.
- ***Stemming:*** It is the process of transforming all the tokens into their stem or root form. Stemming is a swift and straightforward approach that makes the feature extraction process more effortless.

After the prepossessing steps have been completed, the data's polarity and subjectivity can be calculated by labeling the data with the aid of the Textblob method.

Figure 2 shows a pictorial representation of public sentiment during and after lock-down. After labeling the data from TextBlob, figure (a) represents Dataset1 which keeps tweets at the time of lockdown. While Dataset2 shows the post lockdown sentiments.

Algorithm 1: ***Sentiment Analysis of Twitter social media Data (SATD)***

Input:
Twitter social media Data (TD)
Output :
Sentiments of TD

1. Take tweets dataset (TD) by Twint from March to June 2020. All tweets are based upon COVID-19.
2. Preprocess and clean TD by using pandas
3. Find sentiment from TD by using TextBlob *sentiment* function
4. Sentiment_Value = TextBlob(sentence).sentiment[0]
   if Sentiment Value = = 0
   Return 0 //" Neutral"
   elif Sentiment_Value > 0:
   Return 1 //"Positive"
   else:
   Return 2 //"Negative"
5. Write the table with Sentiment_Value
6. Use NLP and machine learning algorithms to find accuracy on COVID 19 TD.

3.3 Proposed approach

This section proposed an approach for the analysis and prediction of Twitter social media data through a machine learning model. It is proposed an algorithm named Sentiment Analysis of Twitter social media Data (SATD). We have taken the dataset from twint, twint is a Twitter social media intelligence tool.

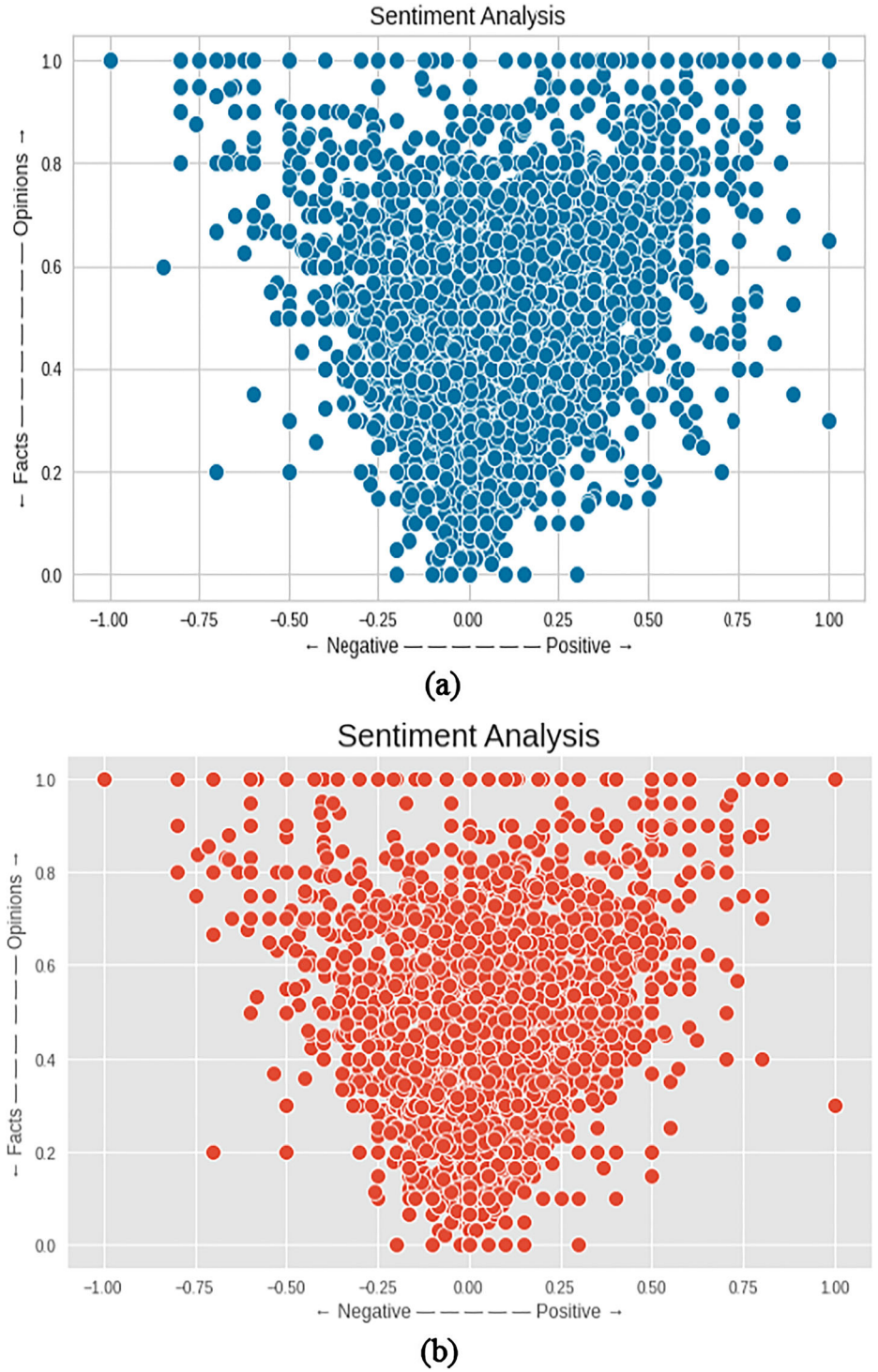


Fig. 2 Sentiment analysis visualization of Dataset1 (**a**) and Dataset2 (**b**)

Here we are analyzing Twitter social media data and finding the sentiment from it. We have categorized and indexed three types of sentiments.

1. Neutral Sentiment- Index 0
2. Positive Sentiment – Index 1
3. Negative Sentiment- Index 2

Five distinct ML models, such as Random Forest Classifier, Multinomial NB Classifier, Logistic Regression, Support Vector Machine, and Decision Tree, have been used and evaluated. These models test and predict the data and the class label also. The algorithm of the proposed SATD is illustrated in algorithm 1.

4 Experimental results

The performance of the proposed algorithms SATD is evaluated in this section. The experiment was performed on 10,924 Tweets during the lockdown and 10,038 tweets are extracted from Twitter using Twins [17]. The data extracted from social media was first preprocessed, followed by labeling (0-neutral, 1-positive, and 2-negative). In which there are two fields containing tweets and their sentiments. We have created a different machine learning classifier model for calculating different values such as precision, recall, F1-score, and support. We have also drawn the confusion matrix with and without normalization.

For the computation of accuracy, recall, f1-score, support, and precision values, Random Forest Classifier, Multinomial NB Classifier, Logistic Regression, Support Vector Machine, and Decision Tree classifiers are used. In Tables 1 and 2 shows the different values with different classifiers. From Table 1, it is observed that the precision, recall, F1-score, support,

Table 1 Parametric values with different classifiers on Dataset1

Class	precision	recall	F1-score	support	Accuracy
Random Forest Classifier					
0.0	0.91	0.95	0.95	879	0.97
1.0	0.93	0.95	0.96	1379	
2.0	0.95	0.96	0.94	473	
Multinomial NB Classifier					
0.0	0.94	0.92	0.95	879	0.98
1.0	0.94	0.99	0.90	1379	
2.0	1.00	0.93	0.97	473	
Logistic Regression Classifier					
0.0	0.94	0.96	0.95	879	0.96
1.0	0.95	0.94	0.92	1379	
2.0	0.95	0.96	0.90	473	
Support Vector Classifier					
0.0	0.93	0.93	0.97	879	0.97
1.0	0.96	0.96	0.92	1379	
2.0	0.94	0.95	0.93	473	
Decision Tree Classifier					
0.0	0.90	0.93	0.91	879	0.98
1.0	0.94	0.92	0.98	1379	
2.0	0.95	0.99	0.99	473	

Table 2 Parametric values with different classifiers on Dataset2

Class	precision	recall	F1-score	support	Accuracy
Random Forest Classifier					
0.0	0.95	0.96	0.98	944	0.94
1.0	0.95	0.93	0.92	766	
2.0	0.95	0.99	0.90	800	
Multinomial NB Classifier					
0.0	0.95	0.97	0.95	944	0.95
1.0	0.92	0.93	0.97	766	
2.0	0.93	0.90	0.99	800	
Logistic Regression Classifier					
0.0	0.92	0.96	0.98	944	0.98
1.0	0.98	0.90	0.94	766	
2.0	0.90	0.99	0.94	800	
Support Vector Classifier					
0.0	0.94	0.96	0.80	944	0.97
1.0	0.99	0.92	0.96	766	
2.0	0.99	0.91	0.95	800	
Decision Tree Classifier					
0.0	0.93	0.99	0.86	944	0.99
1.0	0.99	0.95	0.97	766	
2.0	0.97	0.94	0.96	800	

and accuracy are better for Dataset1. The accuracy obtained for the Random Forest Classifier, Multinomial NB Classifier, Logistic Regression, Support Vector Machine, and Decision Tree classifiers is 0.97, 0.98, 0.96, 0.97, and 0.98 respectively. This is due to the fact the selected classifiers are effectively classifieds based on the sentiments.

Moreover, Dataset2 is also analyzed with the proposed five classifiers and the outcomes are recorded in Table 2. The accuracy of the Random Forest Classifier, Multinomial NB Classifier, Logistic Regression, Support Vector Machine, and Decision Tree classifiers are 0.94, 0.95, 0.98, 0.97, and 0.99 respectively.

4.1 Confusion matrix

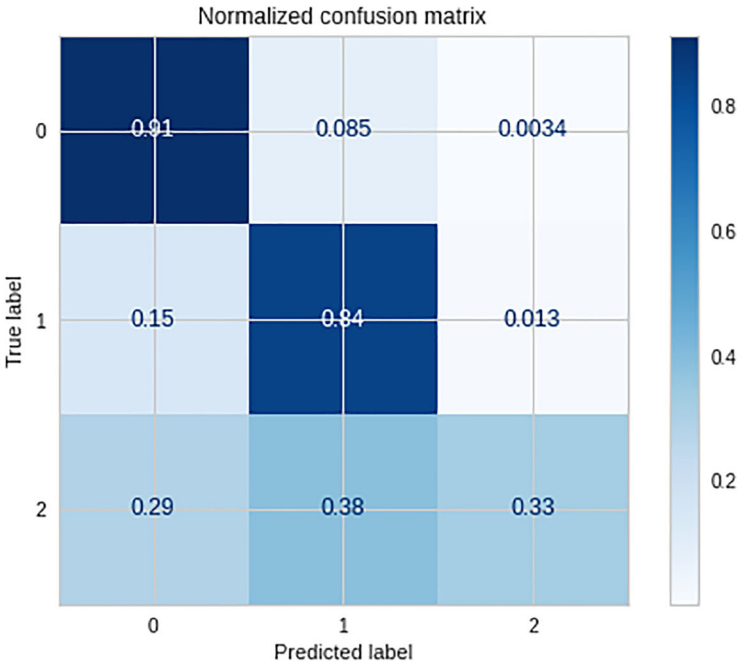
It is a matrix to mention the performance of a classification algorithm. The accuracy itself can be misleading because there are certain reasons for misleading, as an unequal number of observations or more than two classes in your dataset. So overcoming these problems uses a confusion matrix. This matrix shows the values for accurate estimations. This describes the classification model is getting right and what types of errors it is making.

TPR: True Positive Ratio Shows the positive class ratio to be correctly classified.

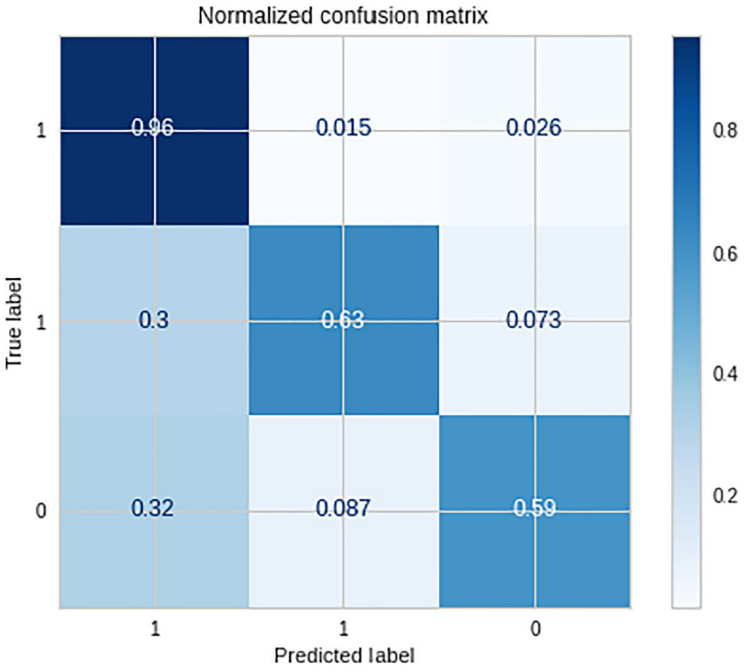
FPR: False positive ratio shows us what amount of the negative class was wrongly identified by the classifier.

4.2 Precision and recall

Precision is the ratio of correctly identified among the total instances obtained. A recall is a fraction of the total number of cases that have been identified.



(a)



(b)

Fig. 3 Confusion matrix using random forest classifier for Dataset1 (**a**) and Dataset2 (**b**)

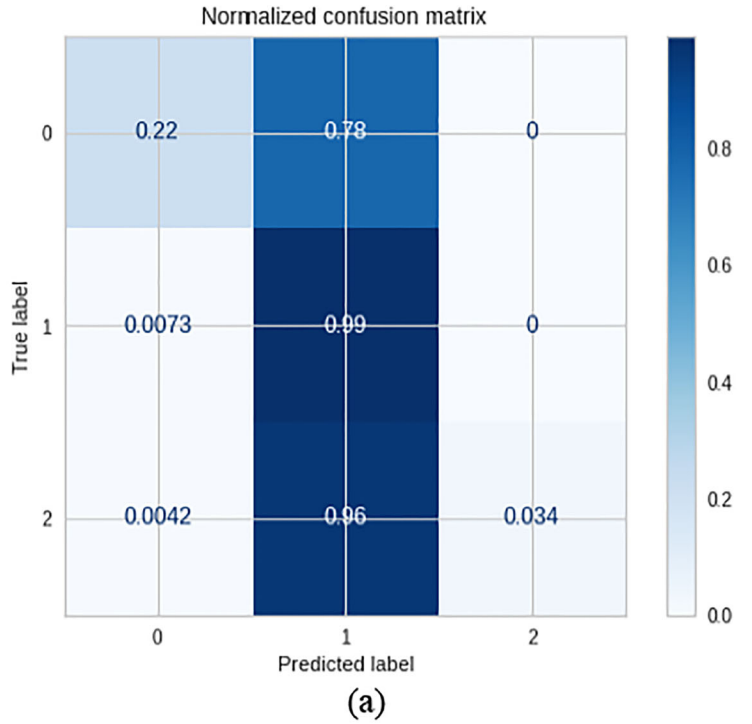
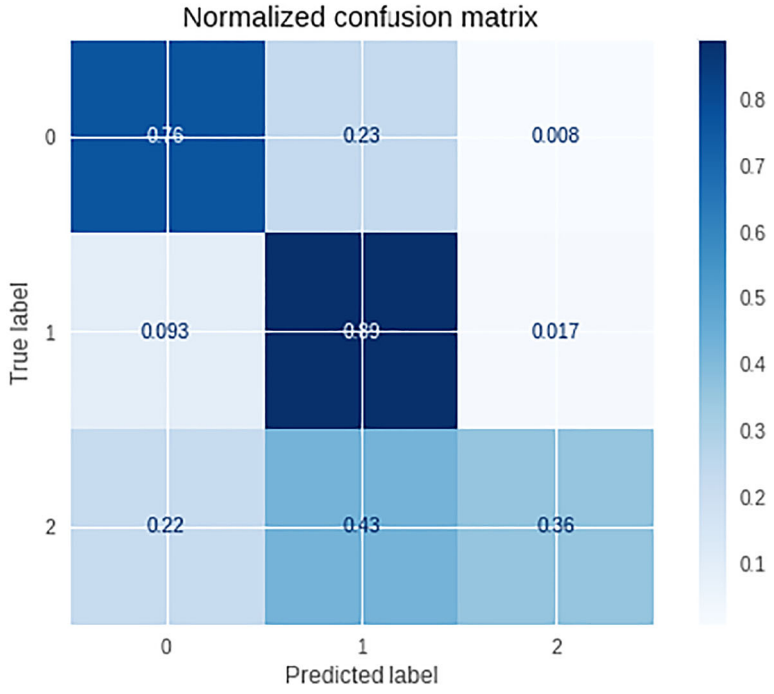
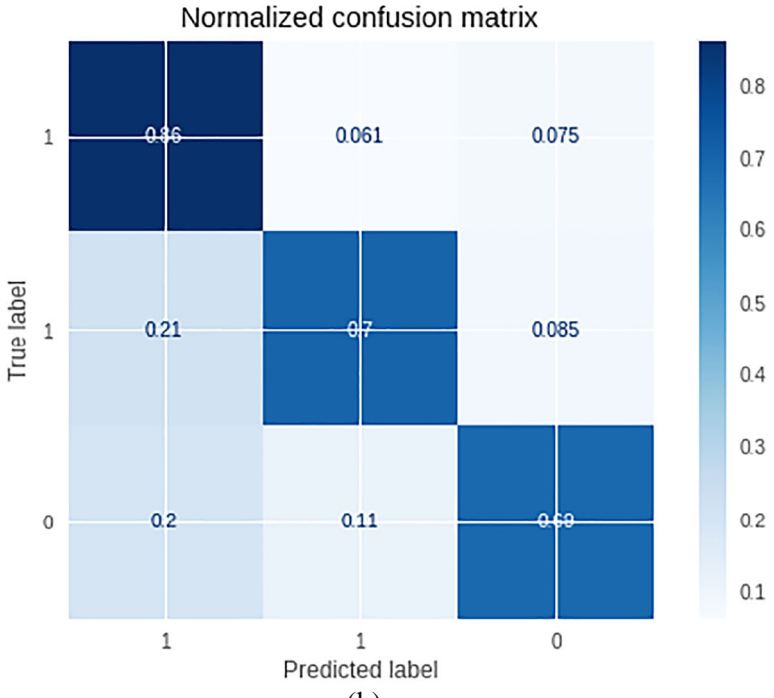


Fig. 4 Confusion matrix using multinomial NB classifier for Dataset1 (**a**) and Dataset2 (**b**)



(a)



(b)

Fig. 5 Confusion matrix using logistic regression classifier for Dataset1 (**a**) and Dataset2 (**b**)

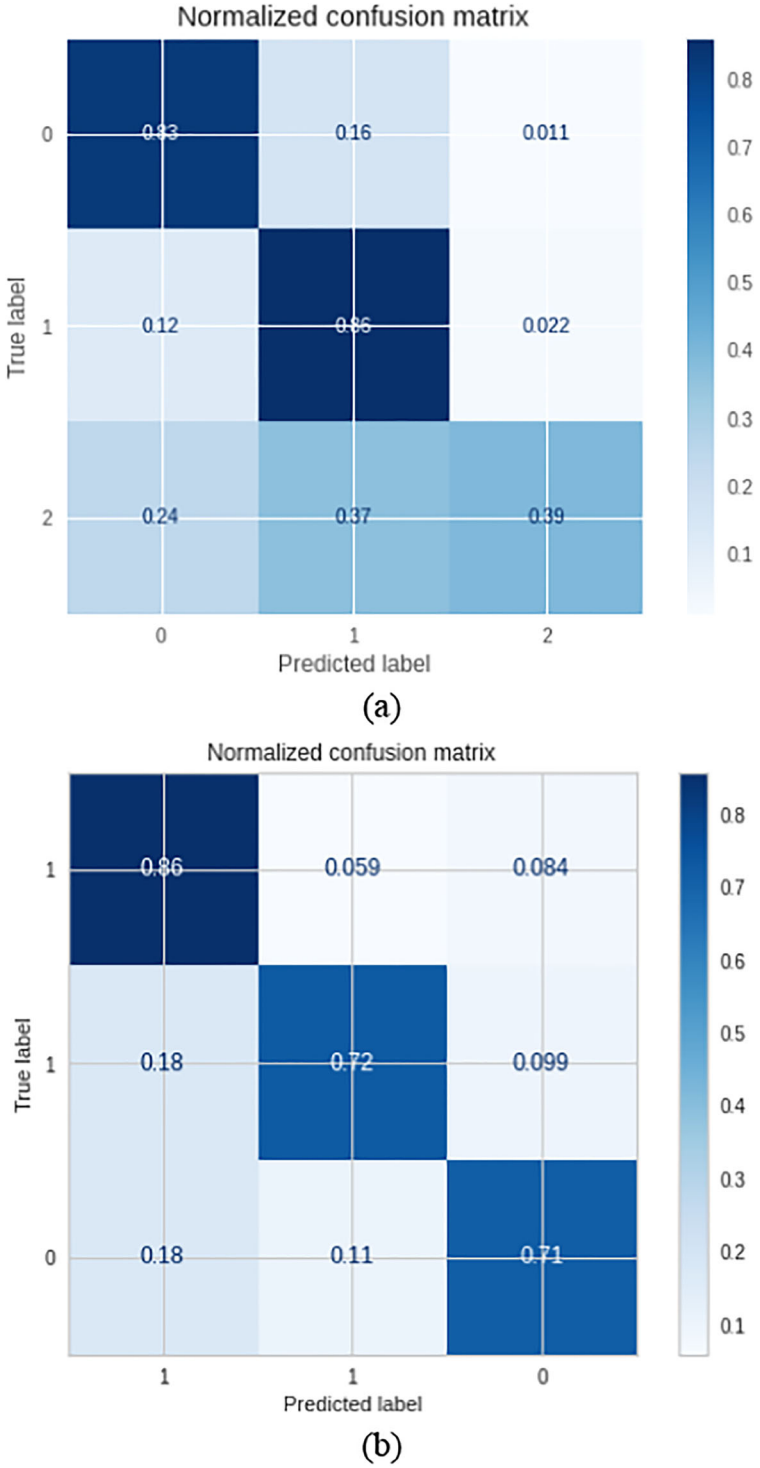
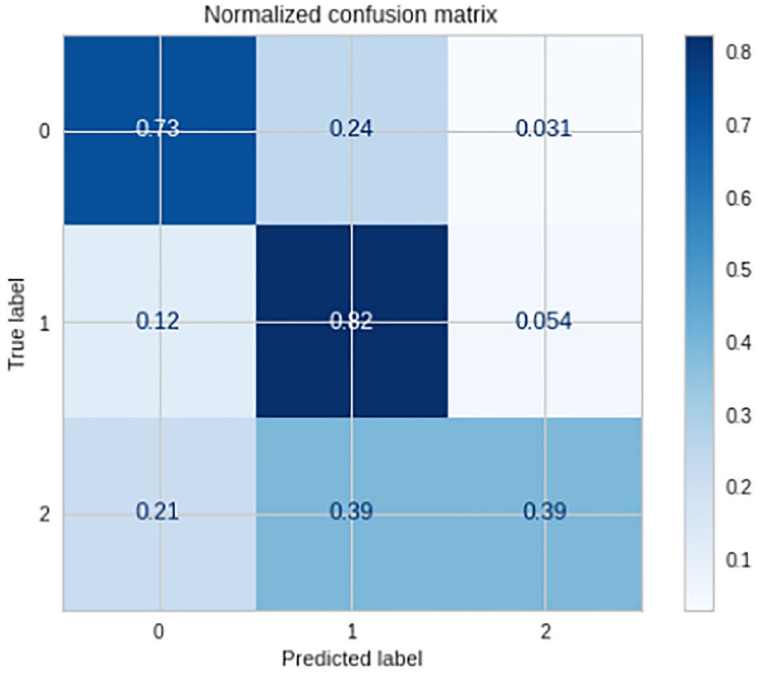


Fig. 6 Confusion matrix using support vector machine for Dataset1 (**a**) and Dataset2 (**b**)



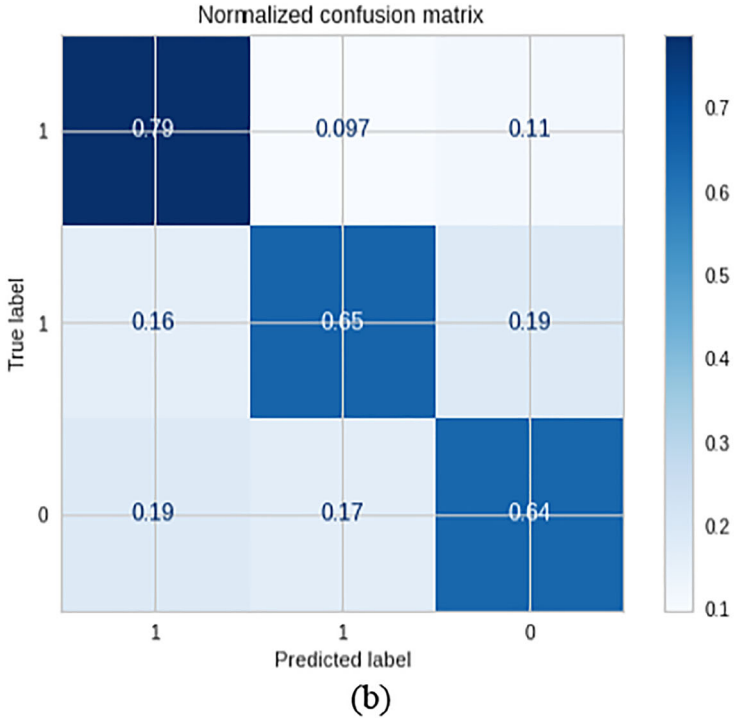
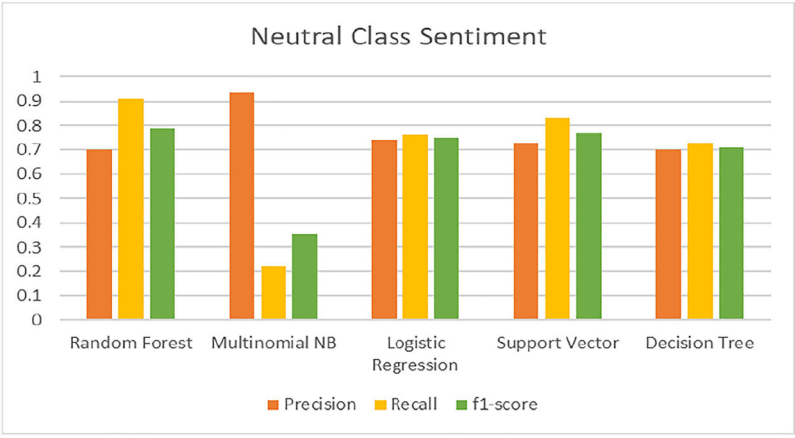
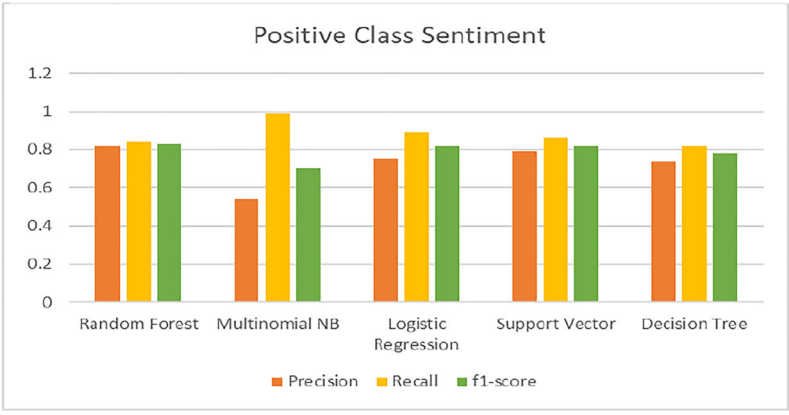


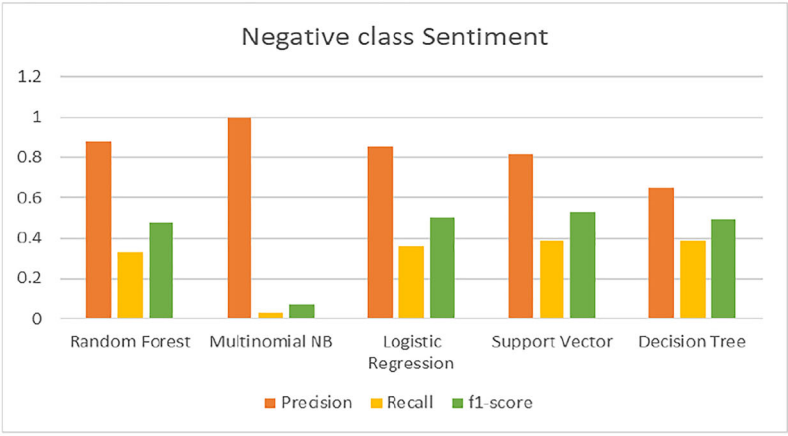
Fig. 7 Confusion matrix using decision tree classifier for Dataset1 (**a**) and Dataset2 (**b**)



(a)

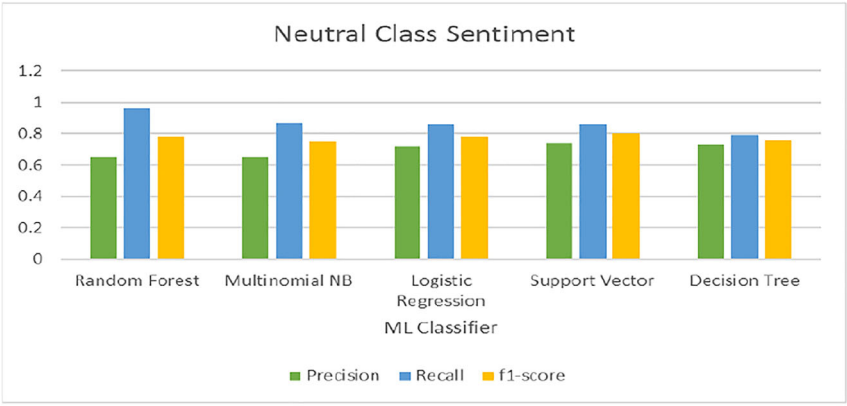


(b)

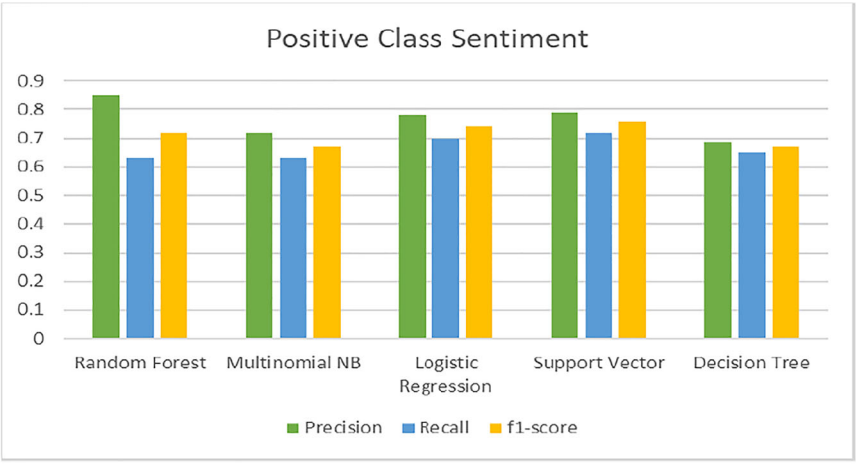


(c)

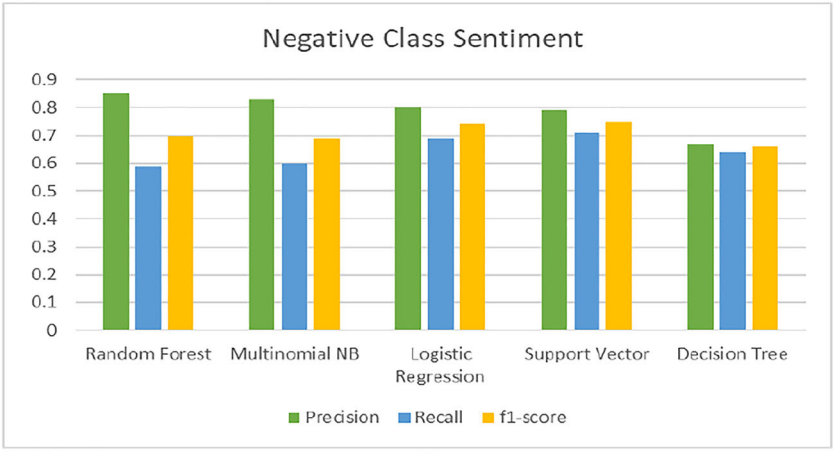
Fig. 8 Precision, Recall and F1-score results (**a**) Neutral, (**b**) Positive, and (**c**) Negative class for Dataset1



(a)



(b)



(c)

Fig. 9 Precision, Recall, and F1-score results (**a**) Neutral, (**b**) Positive, and (**c**) Negative class for Dataset2

4.3 F1-score

It is used for the accuracy of tests and also depends on the precision and recall values. The F1-score can be viewed as the precision and recall weighted average, where the F1-score achieves its highest value at 1 and lowest at 0.

4.4 Receiver operating characteristic (ROC)

It also measures the performance of classification problems. It uses a threshold value for filtration for an outcome. ROC is a probability curve that maps the TPR against FPR at multiple threshold values, that represents the degree or measure of separability. Does this curve show how much the model is capable of distinguishing between classes?

Figure 3, shows the confusion matrix with normalization. We have constructed these matrixes by using a random forest machine learning classifier for both datasets.

Figure 4, shows the confusion matrix without normalization and normalization. We have constructed this matrix by using a multinomial Naïve Bayes machine learning classifier.

Figure 5, shows the confusion matrix without normalization and normalization. We have constructed this matrix by using a logistic regression machine learning classifier.

Figure 6, shows the confusion matrix without normalization and normalization. We have constructed this matrix by using a support vector machine (SVM) machine learning classifier.

Figure 7, shows the confusion matrix without normalization and normalization. We have constructed this matrix by using a decision tree machine learning classifier.

4.5 Discussion

Figure 8, shows the negative, neutral, positive class sentiments of precision, recall, and f1-score values by using five ML classifiers. From the figure, it is evident that the precision value of multinomial NB is high in the case of the Neutral class, and recall of RF is higher than other

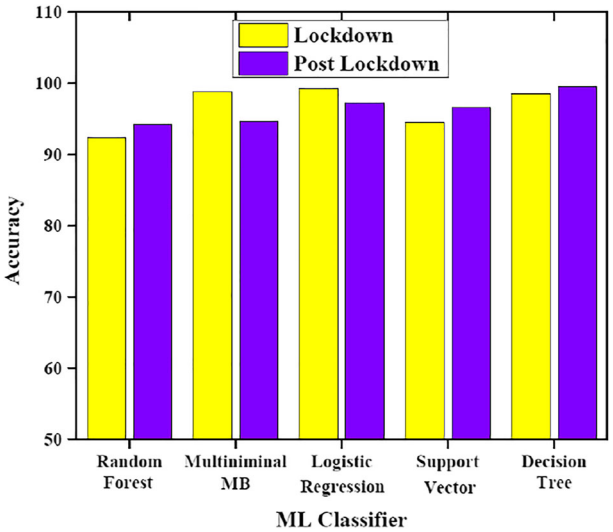


Fig. 10 Measurement of Accuracy for different ML classifiers

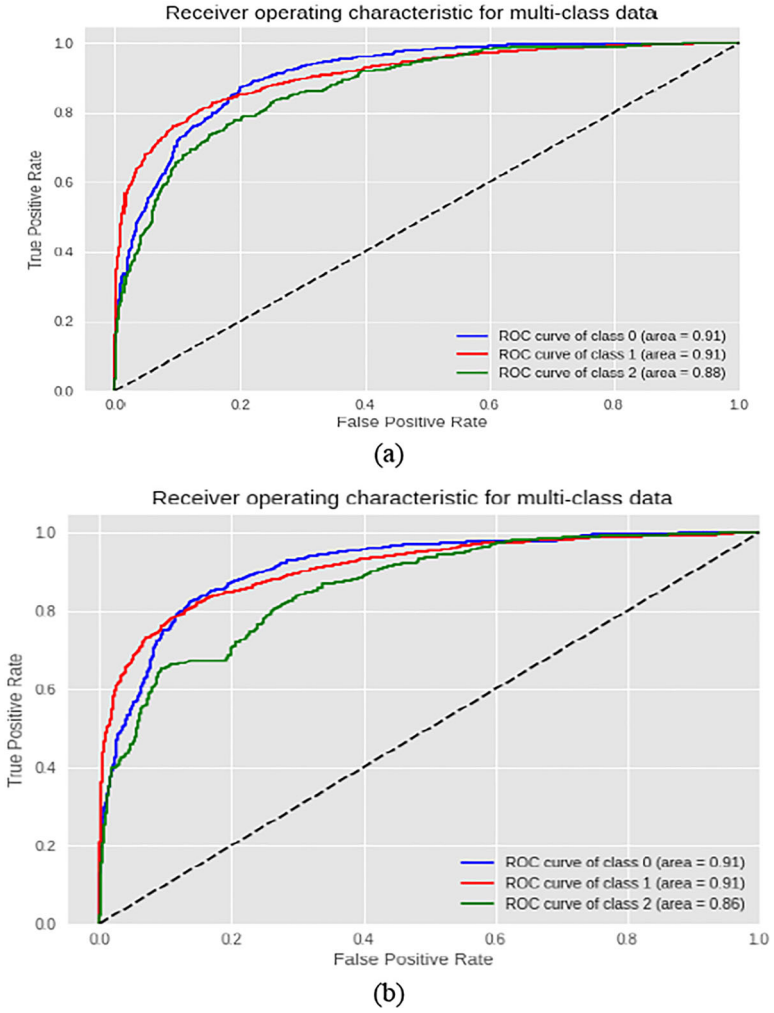


Fig. 11 ROC graphs of logistic regression (**a**) and support vector machine (**b**), classifier during lockdown Tweets

classifiers. In terms of recall, they achieved more or less the same value. In terms of positive class sentiment, the multinomial NB achieves better recall, and the other values are more or less the same. In the case of negative class sentiment, the precision value is higher for the multinomial NB. The analyses have been made for Dataset1.

Figure 9, shows the negative, neutral, positive class sentiments of precision, recall, and F1-score values by using five ML classifiers. For Dataset2 the value of recall is higher for RF classifier for the Neutral class sentiment. For the positive class sentiment too, the RF achieves more than the other classifiers in terms of precision. For the case of negative class sentiment, the precision value of RF is higher than the other classifiers. Meanwhile, the other values are more or less the same for all cases.

Figure 10, shows the accuracy by using five ML classifiers. From the graph, it is clear that the Logistic Regression, MNB shows better results than the others for both lockdown and post lockdown situations. However, the accuracy of the tweet classification by all the five classifiers

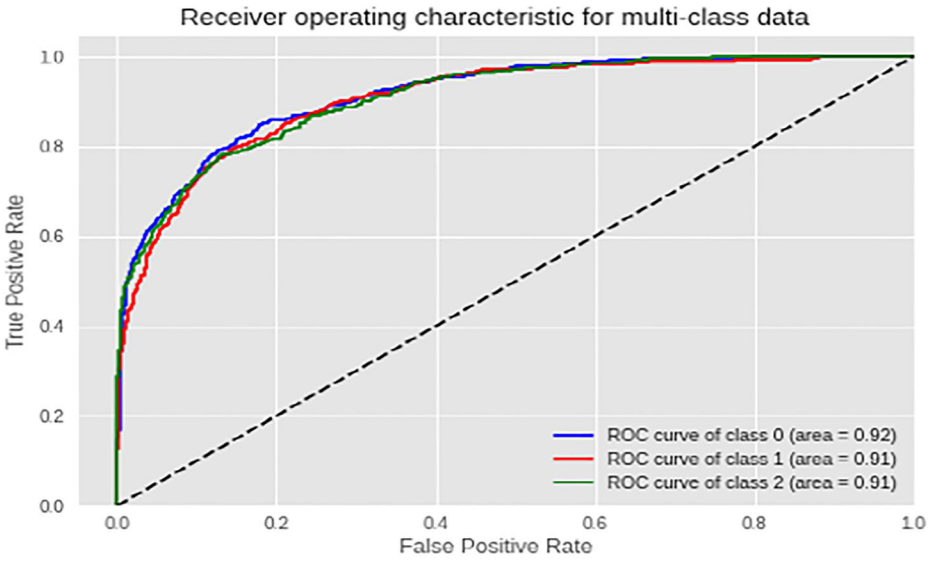


Fig. 12 ROC graphs of logistic regression (**a**) and support vector machine (**b**), classifier Post lockdown Tweets

is higher and provides better results as depicted in the figure. Using this the ROC curve can be plotted which is described in the next section.

In this research, by analyzing two different datasets, we found that people were doing more positive tweets at the lockdown time so that there is no fear. Because in lockdown, people were using social media and television more and more while sitting at home. Positive and neutral tweets were more so that people should support the government in preventing the infection of Corona. After the end of the lockdown, the public's problems, such as employment, public transport, public gathering, educational institutions, etc., the government's strict guidelines kept coming from time to time. Due to which the public had to face many difficulties, due to all these reasons, the sentiment of people was increased in the negative direction.

Figure 11 shows Receiver Operating Characteristic (ROC) graphs. It is clear from the plot that the logistic regression classifier and SVM classifier are well into the neutral and positive classes prediction. Therefore we can say that logistic regression did a better job of classifying the positive class in Dataset1. In Fig. 12 Shows Dataset2, the SVM model classified negative tweets more accurately than the logistic regression model.

5 Conclusion

This paper proposed an algorithm named as Sentiment Analysis of Twitter social media Data (SATD). That algorithm analyzed and predictd the sentiment of tweets based upon COVID-19. The proposed approach predicted the value of precision, recall, and F1-score and support by different machine learning classifiers, such as random forest, multinomial Naïve Bayes, logistic regression, support vector machine, and decision tree on two datasets; first, one extracted from Twitter during the lockdown and second one collected after lockdown. We used ML models programming model with python used pandas library for implementing our algorithm because these types of the library were capable of handling such kinds of extensive data and getting relevant information within limited time and space. The proposed approach predicted the sentiments of Twitter social media data with various parameters. In the future, it can be used sparks technology and deep learning for further processing these data for extracting more accurate results.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Ansari GJ, Shah JH, Yasmin M, Sharif M, Fernandes SL (2018) A novel machine learning approach for scene text extraction. Futur Gener Comput Syst 87:328–340
2. Basha SM, Rajput DS (2017) Evaluating the impact of feature selection on overall performance of sentiment analysis. In: Proceedings of the 2017 international conference on information technology. pp. 96-102

3. Basha SM, Rajput DS (2018) Parsing based sarcasm detection from literal language in tweets. Recent Pat Comput Sci 11(1):62–69
4. Basha SM, Rajput DS (2018) A supervised aspect level sentiment model to predict overall sentiment on tweeter documents. Int J Metadata Semant Ontol 13(1):33–41
5. Basha SM, Rajput DS (2019) An innovative topic-based customer complaints sentiment classification system. Int J Bus Innov Res 20(3):375–391
6. Basha SM, Rajput DS (2019) A roadmap towards implementing parallel aspect level sentiment analysis. Multimed Tools Appl 78(20):29463–29492
7. Dubosson F, Bromuri S, Schumacher M (2016) A python framework for exhaustive machine learning algorithms and features evaluations. In: 2016 IEEE 30th international conference on advanced information networking and applications (AINA). IEEE. pp. 987-993
8. Fu KW, Liang H, Saroha N, Tse ZTH, Ip P, Fung ICH (2016) How people react to Zika virus outbreaks on twitter? A computational content analysis. Am J Infect Control 44(12):1700–1702
9. Gal-Oz N, Gonen Y, Gudes E (2019) Mining meaningful and rare roles from web application usage patterns. Computers & Security 82:296–313
10. Gowthul Alam MM, Baulkani S (2019a) Geometric structure information based multi-objective function to increase fuzzy clustering performance with artificial and real-life data. Soft Comput 23(4):1079–1098
11. Gowthul Alam MM, Baulkani S (2019b) Local and global characteristics-based kernel hybridization to increase optimal support vector machine performance for stock market prediction. Knowl Inf Syst 60(2): 971–1000
12. Gui X, Wang Y, Kou Y, Reynolds TL, Chen Y, Mei Q, Zheng K (2017) Understanding the patterns of health information dissemination on social media during the Zika outbreak. In: AMIA annual symposium proceedings. American medical informatics association. Vol. 2017, p. 820
13. Haseena KS, Anees S, Madheswari N (2014) Power optimization using EPAR protocol in MANET. Int J Innov Sci Eng Technol 6:430–436
14. Hassan BA (2020) CSCF: a chaotic sine cosine firefly algorithm for practical application problems. Neural Comput Applic 33(12):7011–7030
15. Hassan BA, Rashid TA (2020) Datasets on statistical analysis and performance evaluation of backtracking search optimisation algorithm compared with its counterpart algorithms. Data in Brief 28:105046
16. Hassan BA, Rashid TA, Mirjalili S (2021) Formal context reduction in deriving concept hierarchies from corpora using adaptive evolutionary clustering algorithm star. Complex Intell Syst 7(5):2383–2398
17. Cody zacharias(2020) https://pypi.org/project/twint/ (n.d.). Accessed June 2020
18. Ibrahim MNM, Yusoff MZM (2015) Twitter sentiment classification using naive Bayes based on trainer perception. In: 2015 IEEE conference on e-learning, e-management and e-services (IC3e). IEEE. pp. 187–189
19. Ji X, Chun SA, Wei Z, Geller J (2015) Twitter sentiment classification for measuring public health concerns. Soc Netw Anal Min 5(1):13
20. Kavitha D, Ravikumar S (2021) IOT and context-aware learning-based optimal neural network model for real-time health monitoring. Trans Emerg Telecommun Technol 32(1):e4132
21. Kucukyilmaz T, Cambazoglu BB, Aykanat C, Baeza-Yates R (2017) A machine learning approach for result caching in web search engines. Inf Process Manag 53(4):834–850
22. Kuppusamy KS (2018) Machine learning based heterogeneous web advertisements detection using a diverse feature set. Futur Gener Comput Syst 89:68–77
23. Lapuerta P, Azen SP, Labree L (1995) Use of neural networks in predicting the risk of coronary artery disease. Comput Biomed Res 28(1):38–52
24. Li S, Wang Y, Xue J, Zhao N, Zhu T (2020) The impact of COVID-19 epidemic declaration on psychological consequences: a study on active Weibo users. Int J Environ Res Public Health 17(6):2032
25. Li L, Zhang Q, Wang X, Zhang J, Wang T, Gao TL, Wang FY (2020) Characterizing the propagation of situational information in social media during covid-19 epidemic: a case study on weibo. IEEE Trans Comput Soc Syst 7(2):556–562
26. Nair MR, Ramya GR, Sivakumar PB (2017) Usage and analysis of twitter during 2015 Chennai flood towards disaster management. Procedia Comput Sci 115:350–358
27. Nanda C, Dua M, Nanda G (2018) Sentiment analysis of movie reviews in hindi language using machine learning. In: 2018 international conference on communication and signal processing (ICCSP). IEEE. pp. 1069-1072
28. Narendra B, Sai KU, Rajesh G, Hemanth K, Teja MC, Kumar KD (2016) Sentiment analysis on movie reviews: a comparative study of machine learning algorithms and open source technologies. Int J Intell Syst Appl 8(8):66–70
29. Ndaïrou F, Area I, Nieto JJ, Torres DF (2020) Mathematical modeling of COVID-19 transmission dynamics with a case study of Wuhan. Chaos, Solitons Fractals 135:109846

30. Nirmal Kumar SJ, Ravimaran S, Alam MM (2020) An effective non-commutative encryption approach with optimized genetic algorithm for ensuring data protection in cloud computing. Comput Model Eng Sci 125(2):671–697
31. Nisha S, Madheswari AN (2016) Secured authentication for internet voting in corporate companies to prevent phishing attacks. Int J Emerg Technol Comput Sci Electron (IJETCSE) 22(1):45–49
32. Pyrc K, Jebbink MF, Vermeulen-Oost W, Berkhout RJ, Wolthers KC, Wertheim-van PD, Kaandorp J, Spaargaren J, Berkhout B (2004) Identification of a new human coronavirus. Nat Med 10(4):368–373
33. Raghavendra TS, Mohan KG (2019) Web mining and minimization framework design on sentimental analysis for social tweets using machine learning. Procedia Comput Sci 152:230–235
34. Ravikumar S, Kavitha D (2020) IoT based home monitoring system with secure data storage by Keccak–chaotic sequence in cloud server. J Ambient Intell Humaniz Comput 12:1–13
35. Rejeesh MR (2019) Interest point based face recognition using adaptive neuro fuzzy inference system. Multimed Tools Appl 78(16):22691–22710
36. Rustam F, Khalid M, Aslam W, Rupapara V, Mehmood A, Choi GS (2021) A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. PLoS One 16(2):0245909
37. Sahoo K, Samal AK, Pramanik J, Pani SK (2019) Exploratory data analysis using Python. Int J Innov Technol Explor Eng (IJITEE) 8(12):2019
38. Saif H, He Y, Fernandez M, Alani H (2016) Contextual semantics for sentiment analysis of twitter. Inf Process Manag 52(1):5–19
39. Sajib MI, Shargo SM, Hossain MA (2019) Comparison of the efficiency of machine learning algorithms on twitter sentiment analysis of Pathao. In: 2019 22nd international conference on computer and information technology (ICCIT). IEEE. pp. 1-6
40. Sear RF, Velásquez N, Leahy R, Restrepo NJ, El Oud S, Gabriel N, Lupu Y, Johnson NF (2020) Quantifying COVID-19 content in the online health opinion war using machine learning. IEEE Access 8: 91886–91893
41. Singh J, Singh G, Singh R (2017) Optimization of sentiment analysis using machine learning classifiers. Hum-centric Comput Inf Sci 7(1):1–12
42. Soriano CR, Roldan MDG, Cheng C, Oco N (2016) Social media and civic engagement during calamities: the case of twitter use during typhoon Yolanda. Philipp Political Sci J 37(1):6–25
43. Sundararaj V (2019) Optimised denoising scheme via opposition-based self-adaptive learning PSO algorithm for wavelet-based ECG signal noise reduction. Int J Biomed Eng Technol 31(4):325
44. Sundararaj V, Muthukumar S, Kumar RS (2018) An optimal cluster formation based energy efficient dynamic scheduling hybrid MAC protocol for heavy traffic load in wireless sensor networks. Comput Secur 77:277–288
45. Sundararaj V, Anoop V, Dixit P, Arjaria A, Chourasia U, Bhambri P, MR R, Sundararaj R (2020) CCGPA-MPPT: Cauchy preferential crossover-based global pollination algorithm for MPPT in photovoltaic system. Prog Photovolt Res Appl 28(11):1128–1145
46. Team E (2020) The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19)—China, 2020. China CDC weekly 2(8):113–122
47. Tripathy A, Anand A, Rath SK (2017) Document-level sentiment classification using hybrid machine learning approach. Knowl Inf Syst 53(3):805–831
48. Van Lent LG, Sungur H, Kunneman FA, Van De Velde B, Das E (2017) Too far to care? Measuring public attention and fear for Ebola using twitter. J Med Internet Res 19(6):7219
49. Vinu S (2016) An efficient threshold prediction scheme for wavelet based ECG signal noise reduction using variable step size firefly algorithm. Int J Intell Eng Syst 9(3):117–126
50. Zaib NAM, Bazin NEN, Mustaffa NH, Sallehuddin R (2017) Integration of system dynamics with big data using python: an overview. In: 2017 6th ICT international student project conference (ICT-ISPC). IEEE. pp. 1-4
51. Zhang X, Ma R, Wang L (2020) Predicting turning point, duration and attack rate of COVID-19 outbreaks in major Western countries. Chaos, Solitons Fractals 135:109829