Information Sciences Letters

Volume 12
Issue 8 *Aug. 2023*

Article 25

2023

Sentiment Analysis on Twitters Big Data Against the Covid- 19 Pandemic Using Machine Learning Algorithms

Awny Sayed
Computer Science Department, Faculty of Science, Minia University, Minia 61519, Egypt,
mostafa.nazier@gmail.com

Mamdouh M. Gomaa
Computer Science Department, Faculty of Science, Minia University, Minia 61519, Egypt,
mostafa.nazier@gmail.com

Mostafa Medhat Nazier
Computer Science Department, Faculty of Science, Minia University, Minia 61519, Egypt,
mostafa.nazier@gmail.com

Information Sciences Letters
An International Journal

http://dx.doi.org/10.18576/isl/120825

Sentiment Analysis on Twitter's Big Data Against the Covid-19 Pandemic Using Machine Learning Algorithms

*Awny Sayed, Mamdouh M. Gomaa and Mostafa Medhat Nazier**

Computer Science Department, Faculty of Science, Minia University, Minia 61519, Egypt

Received: 27 May 2023, Revised: 13 Jul. 2023, Accepted: 23 Jul. 2023.
Published online: 1 Aug. 2023

Abstract: This paper analyzes users' reactions on Twitter to the COVID-19 pandemic, using machine learning and data mining algorithms to classify tweets according to economic and health fears. A large dataset of tweets is explored, extracted, transformed, loaded, cleansed, and analyzed. The proposed framework improves prediction quality with a proposed dictionary that is used to classify tweets. The study compares four supervised machine learning algorithms and finds that people discuss the pandemic's dangers from economic and health perspectives with equal frequency. The Naive Bayes algorithm achieves the highest percentage of correct predictions.

Keywords: Machine Learning (ML); Big Data (BD); Data Mining (DM); Sentiment Analysis (SA); Naive Bayes (NB); Supported Vector Machine (SVM); Decision Trees (DT); Generalized Linear Model (GLM).

1 Introduction

Corona's disease has now become an epidemic, with a rapid and widespread spread in most countries around the world, and the size of the big data issued by social media platforms necessitates analysis of this data for scientific forecasting of what will happen in the short and long term, as well as people's perceptions of this pandemic and how people will spend this isolation time with their families.

The use of Big Data, Machine Learning, and Data Mining algorithms and techniques in analyzing Big Data exported from a social networking platform like Twitter to analyze the tweets of people around the world about this pandemic represents a treasure that will be useful in studying the impact of this pandemic on people, both economically and socially.

Big Data refers to a variety of different formats of data that are produced from different sources [1]. It is a massive amount of unstructured and structured data [2]. Big data technology contributes to development, governance, search, integration, and analytics services for all data types and formats, from transaction and application data to sensor and machine data to image, geospatial, social media, and more. It is generated by web companies that are used for handling structured or unstructured data.

It has a definition using three v's [2] [3].

1- Volume: several factors contribute to volume growth, including data storage, live streaming, etc.

2- Variety: there are various types of data to be supported.

3- Velocity: processes or files are created or completed at a rapid rate.

Through tweets, users can publicly and privately share their thoughts and feelings about various topics on Twitter [4]. The use of sentiment analysis on tweets has become increasingly popular among marketers and consumers, as it allows consumers and marketers to gain insights about products and analyze the market's behavior. Additionally, with the advancement of machine learning and data mining algorithms, sentiment analysis can also become more accurate [4]. To determine if a text includes positive or negative emotions or communicates feelings about a specific issue, sentiment analysis is utilized.

The main contributions of this paper are:-

- Exploring users' reactions on Twitter about COVID-19's effects.
- Studying, extracting, transforming, loading, cleaning, and analyzing a large data set of people's tweets against the COVID-19 pandemic.

*Corresponding author e-mail: mostafa.nazier@gmail.com

- Classifying people's tweets against the COVID-19 pandemic according to economic and health fears.
- Making a comparative study between four supervised machine learning algorithms for classifying people's tweets against the COVID-19 pandemic. These algorithms are Naive Bayes, Supported Vector Machines, Decision Trees, and the Generalized Linear Model.

This paper is divided into six parts, following the introduction of big data's definitions, problems, social media's big data, and sentiment analysis. The literature review and associated studies are discussed in the second part. Machine learning algorithms are discussed in detail in the third part of this paper. The methodology and study's mechanism, which include the proposed framework, are discussed in detail in the fourth part of the paper. Results from the experiments are addressed in the fifth part, while the sixth part provides the conclusion and recommendation.

2 Related Work

This section presents a literature review of sentiment analysis utilizing machine learning algorithms on big data from social media.

In [5], a statistical investigation of the impact of price and brand on the polarity review was described. The authors compared Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), and Decision Trees (DT) based on the Amazon unlocked mobile phone reviews dataset. Since SVM earned the highest values in all the criteria, including F1 score, precision, accuracy, and recall, they concluded that it was the most complete method for their study.

In [6], NB and SVM on document classification were compared. The authors find that SVM was more accurate than the NB classifier.

Despite the Multinomial Naive classifier's performance in text classification, it was not a fully Bayesian classifier, according to [7]. They recommended a Bayesian Multinomial Nave Bayes classifier as a result, and when they used it, the performance was comparable to that of a standard Multinomial Nave Bayes classifier.

[8] dealt with the classification of social events using a newly proposed ontology. Instead of using a single category to label an event, the writers classified it using tags. As a result, this strategy could give an event several tags and successfully pique user attention. Despite simply applying the Random Forest Classifier during the classification process, the authors conducted classification tests that yielded successful results.

A Random Forest classifier for sentiment analysis was suggested in [9]. The authors developed a method to improve hyperactive parameters like the number of trees used to form the decision forest and the number of features used to randomly choose each tree's depth. The results of this investigation showed that the Random Forest classifier could achieve the greatest outcomes by employing optimal hyperactive parameters (Amazon mobile phone reviews text mining: interesting insights, no date).

[10] analyzed sentiment using social data from Twitter. The key contributions of the authors were parts-of-speech (POS)-specific prior polarity features and the usage of a tree kernel to avoid the requirement for time-consuming feature engineering.

[11] analyzed Twitter social data using a variety of techniques, such as ensemble approaches, machine learning, and dictionary-based approaches.

[12] focused on categorizing sentiment, whether it was positive or negative, rather than neutral. They make use of two manually categorized sets: Twitter status updates and reviews of films from the well-known film portal IMDB (www.imdb.com). They employed three classification algorithms—SVM, K-Nearest Neighbors (KNN), and NB—over a set of features that were retrieved from the texts utilizing the methods and field of natural language processing (NLP).

[13] outlined a method for analyzing huge amounts of data from Twitter using topic-based sentiment analysis, stream analysis, sentiment calendars based on pixel cells, and high-density geo-maps. The authors used interactive geo- and time-based visualizations to conduct a visual study of the Twitter time series. They used the aforementioned methodologies to analyze movie tweets, shopping survey data, amusement park and hotel data, and customer feedback to identify customer interest trends.

[14] centered on the sentiment analysis of Twitter data. The authors surveyed sentiment analysis methods, including lexicon-based methods and machine learning techniques. They suggested utilizing machine learning techniques like SVM, NB, and Max Entropy. They looked into the uses and difficulties of Twitter's sentiment analysis. They agreed that while lexicon-based approaches are more accurate in classifying documents, SVM and NB have the highest accuracy.

A straightforward model called NB allows text classification. In this paradigm, a tweet t is given the class ĉ, where

$$\hat{c} = \underset{c}{\operatorname{argmax}}\, P(c\,|t)$$

$$P(c|t) \alpha P(c) \prod_{i=1}^{n} P\,(f_i\,|c)$$

The i-th feature out of a total of n features is represented by f_i in the formula above. Maximum likelihood estimates can be used to determine P(c) and P (f_i |c). The NB technique employs known priori probability and class conditional probability as a type of module classifier. Calculating the likelihood that document D belongs to class C is the objective [15].

Decision trees serve as supervised classifiers, testing features over the entire data set, and their offspring serve as outcomes. The final classifications of the data points are represented by the nodes at the leaf. Typically, a decision tree is first constructed using data points with known labels, and then the model is applied to the test data. The best test condition or choice is P for each node in the tree. The ideal split is determined using the GINI factor. Where $p(j|t)$ is the relative frequency of class j at node t for a particular node t.

$$GINI(t) = 1 - \sum_j \; [p(j|t)]^2$$

When the DT algorithm is used to classify text, the internal nodes can be labeled with words, the branches can have weight tests applied to them, and the leaf nodes can have congruent class labels applied to them. Documents can be classified by iterating over the query structure from the root to a specific leaf that refers to the classification target [15].

SVM, also referred to as support vector machines, is a non-probabilistic binary linear classifier. Divide the points with $\mathcal{Y}$i = 1 and $\mathcal{Y}$i = −1 for a training set of points ($\mathcal{X}$i, $\mathcal{Y}$i) where y indicates a class and x indicates a feature vector, to get the maximum-margin hyperplane. The hyperplane's equation is as follows.

$$w.\, x - b = 0$$

The margin is maximized, denoted by γ, as follows:

$$\max_{w,\gamma} \gamma \;, \text{s. t.} \; \forall i \;, \gamma \leq \mathcal{Y}i(\mathcal{W} \,.\, \mathcal{X}i + b)$$

Generalized Linear Models (GLMs) are typically traditional linear regression models created for continuous response variables given continuous and/or categorical predictors. ANOVA, ANCOVA, and multiple linear regression are all included [16]. The two forms of this model are the weighted least squares model, which has to have calculated coefficients, and the least squares model, which contains known covariates.

3 Proposed Framework

Prediction quality improves with the proposed framework. It uses big data, machine learning, and data mining algorithms to analyze people's tweets on Twitter about the Corona pandemic. The proposed framework presents six phases. It starts with data preprocessing and the process, which includes data collection. Data preparation and filtering are used for data cleansing, integration, and editing. A suitable file format for the mining tool is created based on the raw data and tested.

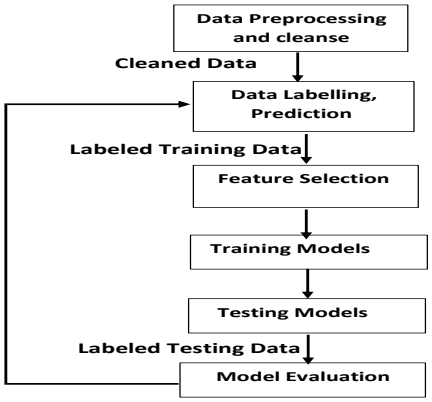

Fig. 1: Overview of the proposed methodology

As shown in Fig. 1, there are five main steps in the big data and machine learning approach.

3.1 Data Preprocessing and Cleanse Process

It is to remove stop words and noise data [15], such as:

- Lower case.
- Remove uniform resource locators (URLs), Hashtags, and targets.
- A sequence of repeated characters is to be handled and the spellings corrected.
- Replace the emotions with their sentiments.
- Remove Non-English tweets.
- Stemming words.

Stemming words is the process by which different word forms are converted into the same canonical form using the stemming algorithm. This step is similar to conflating tokens using the root form, such as computing to compute or connecting to connect [15].

3.2 Data Labeling Prediction Phase

Data labeling is the process of adding labels or tags to data like videos, audio, text, and images. A machine-learning model learns from these labels.

3.3 Feature Selection

One of the most well-known solutions to the high dimensionality of text categorization is feature selection. It is critical to select good features (terms) during text classification. This is a method of improving the categorization's accuracy, effectiveness, and computational efficiency [17].

3.4 Training Models

The process by which the ML algorithm learns from the training data fed to it is known as training models. It assists the model in learning and identifying appropriate values for the attributes involved. This paper employs the classification learning method in these processes, which is a method of learning to categorize unseen data into predefined classes based on a set of training data used for class instance prediction. It employs classification algorithms like SVM, NB, DT, and GLM.

3.5 Testing Models

This paper uses explicit checks to ensure that its model behaves as expected by identifying failure modes.

3.6 Model Evaluation

During this process, the model is evaluated based on its predictive confidence, average accuracy, and overall accuracy. The accuracy of the classification model is measured by predictive confidence. It is calculated by dividing the number of predictions made by the model by the total number of predictions. It has a numerical value between 0 and 1 [18]. The average accuracy of a model is calculated by comparing its predictions to the actual classifications in the test data. Average accuracy is calculated using the following formula [18].

$$\text{Avg. ACC} = (\text{TPs}/(\text{TPs} + \text{FPs}) + \text{TNs}/(\text{FNs} + \text{TNs}))/\text{Number of classes} * 100$$

Where:

TNs is True Negative.

TPs is True Positive.

FNs is False Negative.

FPs is False Positive.

Avg. ACC is Average Accuracy.

The average accuracy achieved for a class of data at a threshold exceeds all other possibilities. In terms of overall

accuracy, the model's prediction accuracy compares with the actual classification in the test data [18]. The formula for calculating overall accuracy is as follows:

$$\text{Over. ACC} = (\text{TPs} + \text{TNs})/(\text{TPs} + \text{FPs} + \text{FNs} + \text{TNs}) * 100$$

Where:

TNs is True Negative.

TPs is True Positive.

FNs is False Negative

FPs is False Positive.

Over. ACC is Overall Accuracy

Performance matrices show how well a model matches actual classification results versus model predictions. By dividing the build data into hold-out samples (the samples created during the split stage of a classification activity), the predictive power of the model can be determined. The desired values have already been determined. The predicted values of the model are compared to the target values. The following functions are carried out by the Performance Matrix [18]:

- Measures whether or not the model predicted correct or incorrect values.
- Indicates how likely the model is to make errors.

Columns represent predicted values, while rows represent actual values. In the upper-right cell of the matrix, for example, the number represents false-positive predictions, or predictions of 1 when the actual value is 0.

Classification models can be evaluated using receiver operating characteristic (ROC) analysis. This type of analysis is only useful for binary classification. ROC curves indicate the degree of discrimination between binary classification models based on their area under the curve. Based on the problem for which the model is being used, the ROC threshold should be set appropriately. [18]. Similarly, ROC curves can be used to compare model results and determine thresholds that provide the highest success rates. The ROC curve is employed to:

- Select thresholds that are highly effective and compare different models.
- Examines the model's ability to forecast positive and negative classes. You can, for example, calculate the probability that the model correctly predicts the positive or negative class.
- Using a classification model, this comparison determines whether the prediction is correct.

Technical steps of Data Mining classification: -

1- Extraction, transferring, and loading of data through the procedural language PL/SQL (PL/SQL) procedure to Oracle tables in Oracle DB XE 18C.

2- Build a dictionary to classify training tweets as below in table 1.

Table 1: data classification dictionary

Target	Words
economical interest	job, economy, unemployment, employment, stock market, exchange, salary, salaries, currency, currencies, market, investment, price, bank, tourism, tourist, tour, museum
health interest	death, deaths, recover, recovered, medicine, drug, drugs, health, serum, plasma, vaccination, vaccine, syndrome, side effects, side effect, symptom, disease, illness, sickness, trouble, complaint, sick, cough, sneeze, temperature, high temperature, tasting, sniff, smell, shortness of breath, breath, pulmonary failure, lung, headache, artificial breathing, ventilator

3- Filter and take only English tweets.

4- Classify training set tweets through a dictionary programmatically using an index built on tweet tables. CREATE INDEX "CRV". "CORONAVIRUS_TWEETS22032020_CAT_IDX" ON "CRV". "CORONAVIRUS_TWEETS22032020" ("TEXT") INDEXTYPE IS "CTXSYS". "CONTEXT"

5- Create an Oracle database (DB) user schema for the Oracle Data Miner (ODM) tool.

6- Create a data mining project in ODM and create the below workflow as shown in Fig. 2.

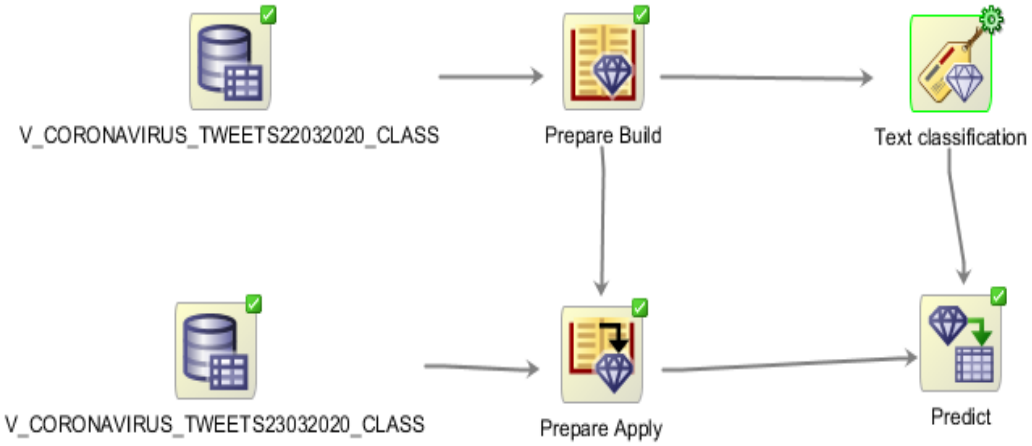

Fig. 2: Text Mining Classification

7- In the workflow, insert data source.

8- Build text transformation in the new column.

9- Build text classification models through data classification algorithms GLM, SVM, DT, and NB and define case ID and target.

10- Define text language as English and stop list.

11- Compare test results through graphs and tables.

12- Add test tweets as the data source needed to be classified.

13- Apply the same transformation to the test data source in the workflow.

14- Then, in a model, add an apply node to instruct the model to predict and classify the test tweets data.

15- After that, view the predicted data.

16- Compare the four algorithms based on predicted classified data.

4 Result and Analysis

The experimental results for the proposed model are presented in this section. Subsection 5.1 describes the datasets used to validate the proposed model's efficiency; subsection 5.2 describes the working environment; and subsections 5.3 and 5.4 explain comparative analysis.

4.1 Datasets Description

As part of the experiment, this paper uses huge datasets. This dataset includes tweets from users who used the mentioned hashtags [19]: #coronavirusoutbreak, #coronavirus, #covid19, #coronavirusPandemic, and #covid_19. It was above 4 million tweets from the start of March 2020 until the end of March, which is the start of the Corona pandemic period. This data set was downloaded from Kaggle. The tweets were from all over the world.

4.2 Working Environment

The Coronavirus tweets were simulated on a local machine equipped with an Intel Core i7 processor, 16 GB of RAM, an NVIDIA Get Force GT 610 (4 GB) GPU, Oracle DB 18C XE, and Oracle Data Miner.

4.3 Classification Results and Performance Evaluation

The authors of this paper examined people's sentiments using tweets extracted from Twitter. These sentiments assisted in gaining an understanding of people's reactions to COVID-19's effects on Twitter. The classification of tweets was the secondary objective of this paper. The data was classified into two categories:

- Economic fears
- Health fears

5% of the whole data set with known predictions was taken for training and testing the models. It is randomly split into 60% for training data and 40% for testing data. As shown in Fig. 3, this is the performance matrix, which is the table where the correct prediction percentage can be calculated based on the values in each column. It shows NB's model has the highest correct predictions, and the decision tree's model has the lowest correct predictions.

Models	Correct Predictions %	Correct Predictions Count	Total Case Count	Total Cost
CLAS_NB_1_5	99.7284	12,852	12,887	0.00000000
CLAS_GLM_1_5	76.4103	9,847	12,887	0.00000000
CLAS_SVM_1_5	96.8806	12,485	12,887	0.00000000
CLAS_DT_1_5	60.0295	7,736	12,887	6,451.40048953

Fig. 3: Performance Matrix of models

After that, classification models were applied to the remaining 95% of the dataset (368,798 tweets), as shown in Fig. 4. Fig. 4 shows the prediction results of the four algorithmic models (NB, DT, SVM, and GLM).

NB_PRED	DT_PRED	SVM_PRED	GLM_PRED	
184024	329119	104494	178446	**economical interest**
184774	39679	264304	190352	**health interest**
368798	368798	368798	368798	**Total**

Fig. 4: prediction results of algorithms 'models

Fig. 4 shows that the NB algorithm predicts that 49.9 percent of tweets have economic interests and 50.1 percent have health interests. The DT algorithm predicts that 89.24 percent of tweets have economic interests and 10.76 percent have health interests. The SVM algorithm predicts that 28.33 percent of tweets have economic interests and 71.67 percent have health interests. The GLM algorithm predicts that 48.39 percent of tweets have economic interests and 51.61 percent have health interests.

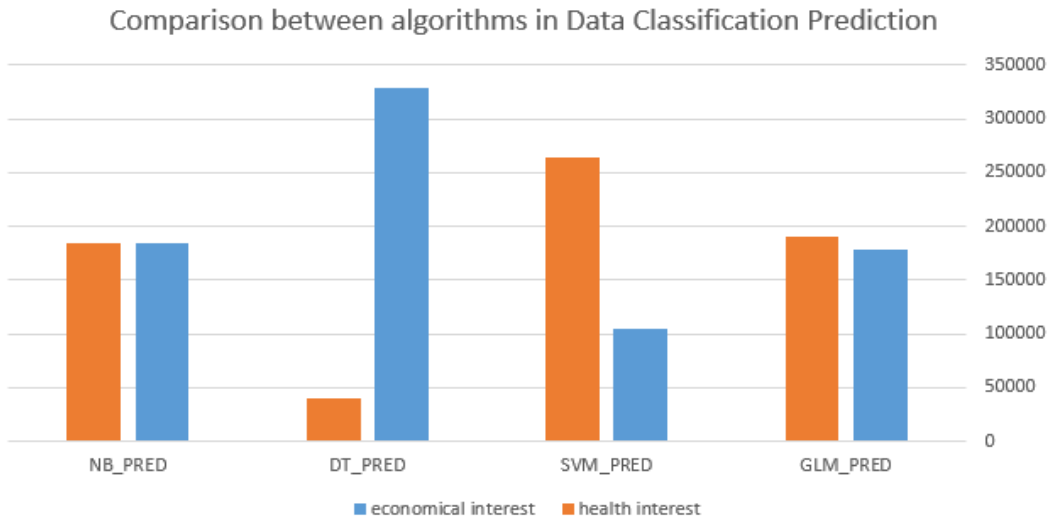

Fig. 5: Comparison between algorithms in data classification's prediction

Fig. 5 shows a comparison between four algorithms for data classification's prediction. It shows that the NB algorithm predicts that the ratio of health interest tweets, with a slight difference, is almost equal to the ratio of economic interest tweets, and the GLM algorithm predicts that the proportion of health interest tweets, with a slight variation, converges

to the proportion of economic interests. However, the figure above also shows that the DT algorithm predicts that the ratio of economic interest tweets has the major number, and the SVM algorithm predicts that the ratio of health interest tweets has the major number.

Fig. 6: All measurements of models

Fig. 6 shows the predictive confidence, overall accuracy, average accuracy, and cost of the four algorithms (NB, SVM, DT, and GLM) in data classification prediction.

4.4 Managerial Implications

The proposed method aims to classify people's tweets according to their economic and health fears through a proposed dictionary. After that, it is used to compare four classification algorithms (NB, SVM, DT, and GLM) to classify coronavirus tweets to investigate people's discussions of the risks of the pandemic from the economic as well as the health directions. This study aims to add to the literature by presenting new findings and recommendations. It helps in the decision-making process and shows how to think about sentiment analysis from different perspectives.

5 Conclusion

This paper presents a classification of people's tweets according to their economic and health fears through a proposed dictionary. It uses four classification algorithms (NB, SVM, DT, and GLM) to classify coronavirus tweets. It demonstrates that people are discussing the risks of a pandemic from both an economic and a health standpoint. Regardless of the media, all governments advised people to stay at home at the start of the pandemic to control the spread of the COVID-19 virus. It also shows that the Naïve Bayes algorithm has the highest percentage of correct predictions in classification. It also shows that the Decision Tree algorithm has the lowest percentage of correct predictions. Naïve Bayes and Supported Vector Machines are more accurate to be used in sentimental analysis. In our future work, we will use deep learning algorithms like convolutional neural networks (CNN) and recurrent and recursive neural networks (RNN) in sentiment analysis on Twitter's Big Data to have longer training times and higher accuracy.

Funding Statement

The authors received no specific funding for this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. Alam and K. A. Shakil, Big data analytics in cloud environment using Hadoop, *arXiv preprint arXiv:1610.04572*, (2016).

[2] H. Venkatesh, S. D. Perur and N. Jalihal, A study on use of big data in cloud computing environment, *International Journal of Computer Science and Information Technologies (IJCSIT)*, 6 (3), 2076-2078 (2015).

[3] S. A. El-Seoud, H.F. El-Sofany, M. Abdelfattah and R. Mohamed, Big data and cloud computing: trends and challenges, *International Journal of Interactive Mobile Technologies*, 11(2), 34–52 (2017).

[4] P. Singh, Y. K. Dwivedi, K. S. Kahlon, A. Pathania and R. S. Sawhney, Can twitter analytics predict election outcome? An insight from 2017 Punjab assembly elections, *Government Information Quarterly*, 37 (2), 101444 (2020), DOI: 10.1016/j.giq.2019.101444

[5] M. Guia, R. R. Silva and J. Bernardino, Comparison of naïve bayes, support vector machine, decision trees and random forest on sentiment analysis, *In Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2019)*, 11, Vienna, Austria, 525-531 (2019).

[6] Z. H. Moe, T. San, M. M. Khin and H. M. Tin, Comparison of naive bayes and support vector machine classifiers on document classification, *in 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)*, IEEE, Nara, Japan, 466–467 (2018), DOI: 10.1109/GCCE.2018.8574785.

[7] S. Xu, Y. Li and W. Zheng, Bayesian multinomial naïve bayes classifier to text classification, *In Advanced Multimedia and Ubiquitous Engineering: MUE/FutureTech*, *Springer*, 11, Seoul, South Korea, 347-352 (2017), DOI: 10.1007/978-981-10-5041-1_57.

[8] M. Rodrigues, R. R. Silva and J. Bernardino, Linking open descriptions of social events (LODSE): A new ontology for social event classification, *Information (Switzerland)*, 9(7), 164 (2018), DOI: 10.3390/info9070164.

[9] H. Parmar, S. Bhanderi and G. Shah, Sentiment mining of movie reviews using random forest with tuned hyperparameters, *International Conference on Information Science*, Kerala, India, 1- 6 (2014).

[10] A. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. J. Passonneau, Sentiment analysis of twitter data, *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, Portland, Oregon, USA, 30-38 (2011).

[11] A. A lsaeedi and M. Z. Khan, A study on sentiment analysis techniques of Twitter data, *International Journal of Advanced Computer Science and Applications*, 10 (2), 361-374 (2019).

[12] K. Bar, Sentiment analysis of movie reviews and twitter statuses, *Machine Learning–Final Project* ,1-12 (2013).

[13] M. Hao, C. Rohrdantz, H. Janetzko, U. Dayal, D. A. Keim et.al., Visual sentiment analysis on twitter data streams, *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, IEEE, Providence, RI, USA, 277-278 (2011).

[14] V. Kharde and P. Sonawane, Sentiment analysis of twitter data: a survey of techniques, *arXiv preprint arXiv*:1601.06971 (2016).

[15] V. Korde and C. N. Mahender, "Text classification and classifiers: A survey," *International Journal of Artificial Intelligence & Applications*, 3 (2), 85-99 (2012).

[16] Penn State Eberly College of Science, Analysis of Discrete Data (STAT 504). https://online.stat.psu.edu/stat504/book/ accessed on 19-06-2021.

[17] B. Harish and M. Revanasiddappa, A comprehensive survey on various feature selection methods to categorize text documents, *International Journal of Computer Applications* , *164* (8), 1-7 (2017).

[18] Sql developer documentation release 4.1, Data Miner User's Guide, accessed on 20-05-2021. Available online at https://docs.oracle.com/cd/E55747_01/doc.41/e58114/test.htm#DMRUG816.

[19] Coronavirus (covid19) Tweets, Tweets using hashtags associated with Coronavirus, accessed on from 18-03-2020 to 20-04-2020.Available online at https://www.kaggle.com/smid80/coronavirus-covid19-tweets

[20] F.B. Hamzah, C. Lau, H. Nazri, D.V. Ligot, G. Lee et al., Corona Tracker: worldwide COVID-19 outbreak data analysis and prediction, Bull World Health Organ, 1(32), 1-32 (2020).

[21] P. Singh, K. S. Kahlon, R. S. Sawhney, R. Vohra and S. Kaur, "Social media buzz created by# nanotechnology: insights from Twitter analytics." *Nanotechnology Reviews* 7, no. 6, 521-528, 2018.

[22] E.-S. M. El-kenawy, A. Ibrahim, N. Bailek, B. Kada, M. Hassan et al., Sunshine duration measurements and predictions in Saharan Algeria region: An improved ensemble learning approach, *Theoretical and Applied Climatology*, 147, 1015–1031 (2022).

[23] A. E. Takieldeen, E. M. El-kenawy, M. Hadwan and R. M. Zaki, Dipper throated optimization algorithm for unconstrained function and feature selection, *Computers, Materials & Continua*, 72 (1), 1465–1481 (2022).

[24] B. Chilwal and A. K. Mishra, Extraction of depression symptoms from social networks, *The Smart Cyber Ecosystem for Sustainable Development*, 307–321 (2021), doi: 10.1002/9781119761655.CH17.

[25] V. Arya, A. K. M. Mishra, and A. González-Briones, Analysis of sentiments on the onset of Covid-19 using Machine Learning Techniques, *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, (2022).

[26] N. Öztürk and S. Ayvaz, Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis, *Telematics and Informatics*, 35 (1), 136-147 (2018).

Appendix

This dataset contains the Tweets of users who have applied the following hashtags:

#covid_19	#coronavirusoutbreak	#coronavirusPandemic	#covid19	#coronavirus