Evaluating Sentiment Classification to Specify Polarity by Lexicon-Based and Machine Learning Approaches for COVID-19 Twitter Data Sets

A. Sathya[1]*, Dr. M.S Mythili[2]

[1]Research Scholar, Department of Computer Science, Bishop Heber College (Autonomous), Affiliated to Bharathidasan University, Trichy – 620017, Tamilnadu, India.

[2]Associate Professor, P.G. Department of Computer Applications, Bishop Heber College (Autonomous), Affiliated to Bharathidasan University, Trichy- 620017, Tamilnadu, India.

Corresponding author E-mail: sathya.cs.res@bhc.edu.in

Abstract

As part of data science, sentiment analysis (SA) applied to social media data is a trending research topic. Identifying positive, negative, or neutral opinions or feelings in the text is the attention of sentiment analysis. In the past few years, Social media platforms have become increasingly popular. In this research, natural language processing (NLP) will be employed to extract useful data and information from unstructured text. .The two methods for sentiment analysis covered in this research are the machine-learning method and the lexicon-based method. The paper examines several lexicon approaches to demonstrate the sentiments from Twitter. To increase classification accuracy, it is necessary to use a reliable method with the highest performance. In this study, classifiers such as Support Vector Machine (SVM) and Naive Bayes (NB) were used together with techniques such as TF-IDF (Term Frequency-Inverse Document Frequency) and BOW (Bag of Words). Each algorithm produces a unique outcome. In order to measure the accuracy of classification, metrics such as Precision, Recall, and F-score are considered. This research shows Support Vector Machine (SVM) with TF-IDF is better than other classifiers with an accuracy of 88%.

Keywords: Sentiment Analysis, Machine Learning Algorithm, Lexicon Based Approach, Social Media.

Introduction

Using data from Covid data sets, natural language processing can be used to extract information about public opinion. The use of sentiment analysis is extensive and effective. In sentiment classifiers, a machine learning technique can be either supervised or unsupervised, and there are pros and cons to each. . The classifier needs both the target data and the labeled training data for the trained approach. In the current work, both supervised and unsupervised approaches are used to identify sentiment [1]. The polarity of the sentiment of any study problem is determined using the supervised learning techniques SVM and Naive Bayes classifiers. The Naive Bayes algorithm finds the probabilistic classifier of the set of assumptions data sets. The Support Vector Machine is a frontier that separates classes of a hyperplane. There is a variety of pre-processing techniques, including tokenization, stop word removal, and hashtag removal, for text normalization. Sentiment analysis is a method of identifying the text's related polarity. It divides importance into three categories, including positive, negative, and neutral significance. The training set performance, along with the correct text value, are needed for sentiment analysis. Numerous studies have been conducted in the area of sentiment analysis, mostly on social media platforms and marketing levels [2]. There are two categories in which sentiment analysis methodology based on dictionary matching and machine learning is required. The term "lexicon" refers to procedures that use dictionaries to analyse texts or words for polarity [3]. Based on the research problem, machine learning techniques use a variety of classifiers, including Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), and Decision Tree (DT). Utilizing popular feature extraction methods including BOW, TF-IDF, and Count-Vectorizer that are appropriate for the study, the specific data were extracted. TF-IDF techniques outperform binary bags of words, which treat all words equally [4]. As a result, sentiment analysis/opinion mining analyses the opinions, evaluations, sentiments, attitudes, appraisals, and emotions of people about products, organizations, services, issues, individuals, topics, events, etc.

Related Works

This section discussed the studies related to sentiment analysis using machine-learning techniques and lexicon-based approaches.

Utilizing Twitter messages, Vishal A. Kharde et al. [9] addressed lexical techniques and machine learning methods. Probabilistic classifiers like"NB and SVM" were utilized for improved outcomes. Machine learning techniques were employed by Rajeswara Rao et al. [16] to forecast election outcomes. To split up long text into smaller units, tokenization was used as a strategy. Support vector machines, in comparison to other classifiers, proved to be more accurate. Rupinder Kaur et al. [17] concentrated on Twitter members' comments. In order to make the data more uniform, text pre-processing was employed to eliminate pointless words and symbols. For future sentiment analysis improvements, the author advised establishing a word dictionary. Lexicon and machine methods were discussed by Reena G. Bhati [18] using various dataset categories. With the use of the Natural Processing approach, text data may be divided into different emotional categories for easier analysis. T Nikil Prakash [19] discussed the concept of sentiment acronyms, emotions, and contextual sentiment. The proposed work is to improve the accuracy and efficiency of existing work. The author focuses on the lexicon method and the data are collected from Twitter comments using the Kaggle website. Numerous parameters including precision, recall, and the F1-score be present to improve the accuracy level. This section discusses works in the area of sentiment analysis that use lexicon-based approaches and machine learning systems.

Methodology

The following is a detailed explanation of how a lexicon- and machine-learning approach was used to discuss how people felt during the Pandemic. In this research paper, the main aim is to obtain a better understanding of the social opinions and persceptives on COVID-19 and how it has changes people's thinking over the past few years. Social media such as t witter is mainly beneficial to extract information related to the user's sentiment, opinions and insights on a numerous number of topics. Twitter is blogging social media platform and it has a huge and growing number of users every day. Most of the twitter users are between the age group of 30 to 65. In twitter, user post short messages or blogs of 140 or fewer characters to tweets about their opinion on coronavirus, to share information and to have talks with their followers. As a result, tweets gathered from Twitter data for sentiment analysis of people's opinions on the coronavirus using machine learning algorithms will aid in analysing user sentiment into three categories:

positive, negative, and negative during the illness epidemic. As shown in the figure below, progressive steps were taken to analyse a specific Twitter comment, particularly from English Tweets. The first step is to gather data from Twitter, which is then pre-processed so that it can be easily read by machines. By using well-known feature extraction techniques like TF-IDF and Bow, specific information is extracted. By using a Lexicon-based approach and Machine learning techniques, it is possible to determine the polarity of a sentiment. Finally, the data are visualized through some graphical representations.

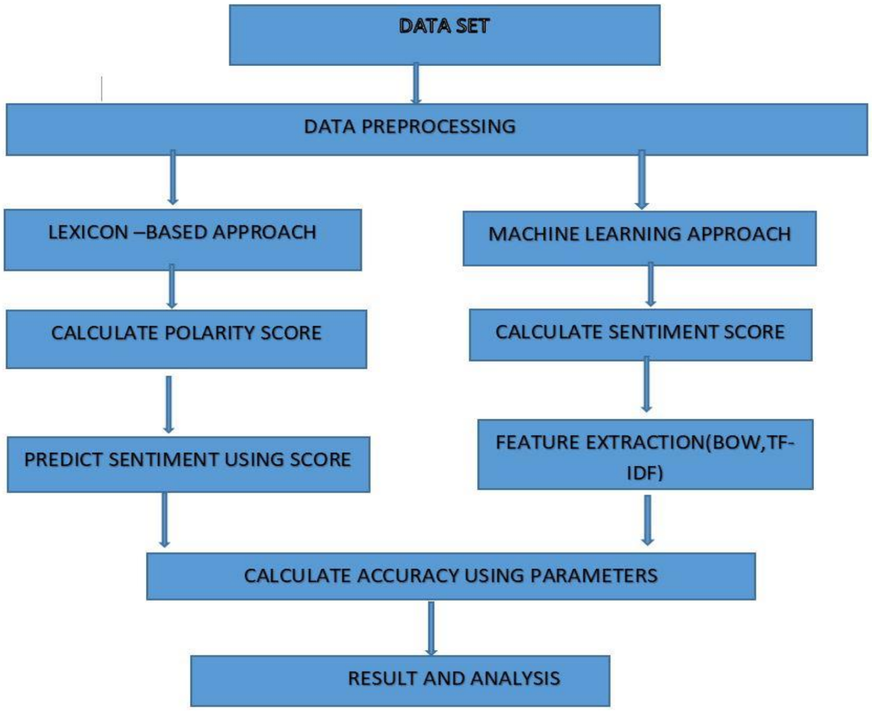

DATA SET
DATA PREPROCESSING
LEXICON –BASED APPROACH
MACHINE LEARNING APPROACH
CALCULATE POLARITY SCORE
CALCULATE SENTIMENT SCORE
PREDICT SENTIMENT USING SCORE
FEATURE EXTRACTION(BOW,TF-IDF)
CALCULATE ACCURACY USING PARAMETERS
RESULT AND ANALYSIS

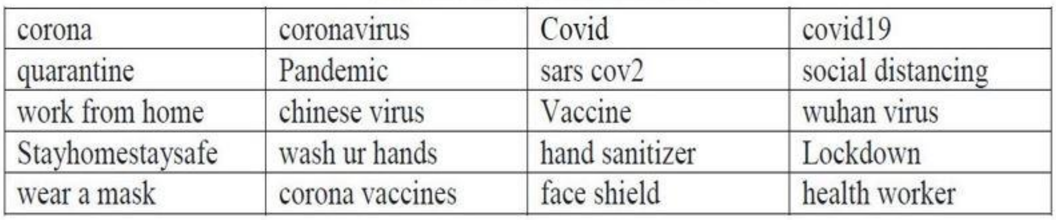
FIGURE 1. Flow Diagram of the proposed model

Data Extraction

Research studies show that 90% of people use Twitter as an electronic database, with the other 10% using online media, such as Reddit, Yelp, Health grades, WeChat profiles, etc. Twitter is considered to be the most well-liked social networking site, with 81.47 million registered members [21]. Tweets, or communications, are exchanged concerning national and worldwide happenings. A total of 200 billion tweets are published annually, or 500 million every day [22]. The tweets are divided into categories based on their topics, which include politics, individual viewpoints, and issues with the country's economy, particularly the COVID-19 epidemic [23].

Using the Twitter API, data is taken from Twitter. A Significant amount of data are being generated by individuals on Twitter. The average Twitter user produces 12 gigabytes of data per day. The service is popular among many people in addition to expressing their opinions on current topics like politics, business, government, and health care [5].

TABLE 1: Collection of Tweets from Twitter



corona	coronavirus	Covid	covid19
quarantine	Pandemic	sars cov2	social distancing
work from home	chinese virus	Vaccine	wuhan virus
Stayhomestaysafe	wash ur hands	hand sanitizer	Lockdown
wear a mask	corona vaccines	face shield	health worker

Data Pre-Processing

During preprocessing, all unnecessary data should be removed from the data in order to improve its quality. Analysis cannot be done on tweets in their raw form.

Effective pre-processing is essential for more accurate analytical results. This process deals with cleaning and preparing text data. Tokenization, which breaks up the text into smaller chunks for each word to be more easily accessed, stop word removal, which gets rid of similar words that don't provide pertinent information, and further processing like stemming and lemmatization may also be used [24]

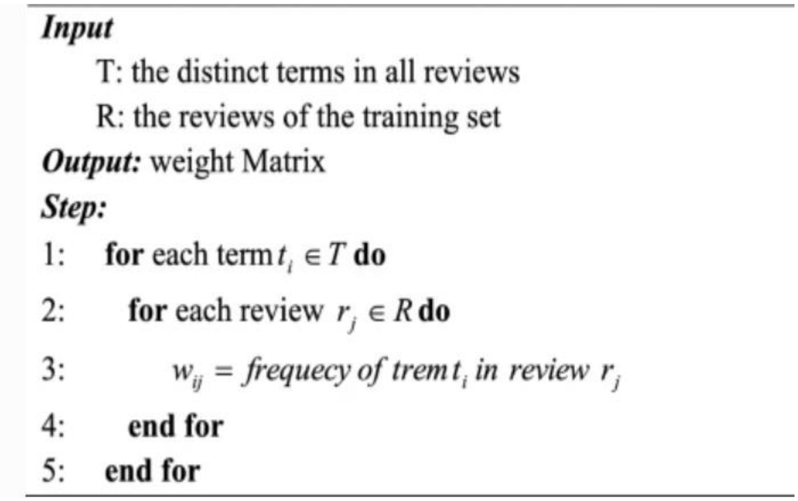
- The data can be processed by deleting extraneous information, including HTML elements, which frequently exist in our text but provide little in the way of sentiment analysis. Therefore, before extracting features, we must make sure that they are eliminated
- Removing numeric and special characters
- Tokenization, which breaks down phrases into words
- The most frequent words in a text that don't contribute anything useful are stopped words

- Stemming is a method that removes the last few characters from a word often leading to incorrect meanings and spelling.

Determining the Sentiment by Calculating the Polarity Score:

The polarity of the tweet is determined by utilising the TF-IDF to determine the phrase weight. Based on how frequently a phrase appears in a dataset of tweets, its positivity and negativity are estimated.

The general algorithm operates as follows:



```
Input
    T: the distinct terms in all reviews
    R: the reviews of the training set
Output: weight Matrix
Step:
for each term $t_i \in T$ do
    for each review $r_j \in R$ do
        $w_{ij} = frequecy\ of\ trem\ t_i\ in\ review\ r_j$
    end for
end for
```

Algorithm 1 : Term frequency-Inverse Document Matrix

From the above algorithm it is clear to determine each word's rank, the TF-IDF is applied to all the words in the dataset. A term with a high TF-IDF rank is related to the supplied tweet and significantly affects the tweet's polarity

Feature Extraction Techniques:

Bag of Words (BOW): "Bag of Words (BOW)" is the best straightforward approach for mining text features is BOW. It will explain where specific terms appear in a document. The vocabulary of words used to create each sentence vector is represented by the bag. This method's performance is assessed, which generally performs better [14].

Term Frequency-Inverse Document Frequency (TF-IDF): Frequency of terms-Inverse Document frequency when a word's TF, or frequency, is counted in a tweet.

$$TFIDF\ score\ for\ term\ i\ in\ document\ j = TF(i,j) * IDF(i)$$

$$where$$

$$IDF = Inverse\ Document\ Frequency$$

$$TF = Term\ Frequency$$

$$TF(i,j) = \frac{\text{Term i frequency in document j}}{\text{Total words in document j}}$$

$$IDF(i) = \log_2\left(\frac{\text{Total documents}}{\text{documents with term i}}\right)$$

$$and$$

$$t = Term$$

$$j = Document$$

By dividing the final TF-IDF score by the combined TF-IDF score for all tweets. Words are developed that are also utilized to encrypt unbreakable text. A statistic called TF-IDF counts how often certain terms appear in different tweets [13].

***Vader*:** Vader (“Valence Aware Dictionary and Sentiment Reasoner”) is a lexical function called VADER analyses text sentimentality, paying particular attention to feelings expressed on Twitter [15]

***Text blob*:** To polarize tweets, text blob objects were used. It was possible to determine the polarity value of each individual tweet on the basis of the Covid data. Using the specific properties, it determines the polarity value of the item. Happy values corresponded to positive emotions, while negative values corresponded to negative sentiments, and polarities that were close to zero were classified as neutral sentiments[15].From Figure(2) the approaches of sentiment analysis is clearly explained.

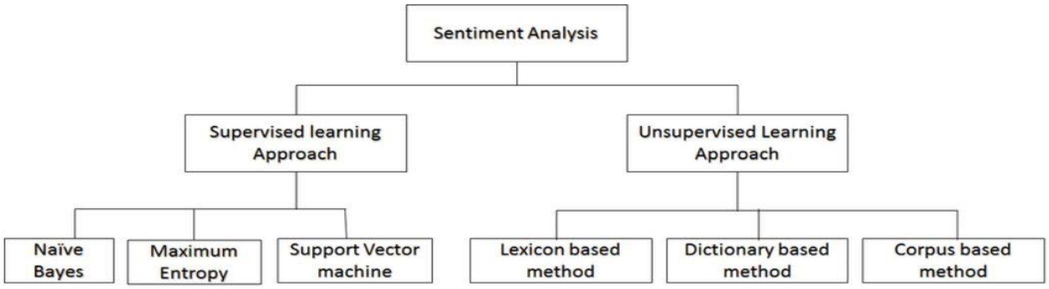


FIGURE 2. Approaches of Sentiment Analysis

Machine Learning Classifiers:

Naive Bayes: Using the Bayes theorem, Naive Bayes is a traditional classification algorithms. Alternatively, it is referred as independent Bayes. It has been highly suggested for classification [11] because of its speed and simplicity. According to the formula below, the probability is calculated as follows:

$$\underbrace{P(c \mid x)}_{\text{Posterior Probability}} = \frac{\overbrace{P(x \mid c)}^{\text{Likelihood}}\,\overbrace{P(c)}^{\text{Class Prior Probability}}}{\underbrace{P(x)}_{\text{Predictor Prior Probability}}}$$

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Support Vector Machine: One of the best techniques for categorising text on the hyperplane is SVM. From the fig (3) the term "input and output format" is used to describe the hyperplane of X and Y labels. Either a good or negative result could be obtained. SVM classifiers are costlier and will execute significantly [12][20].

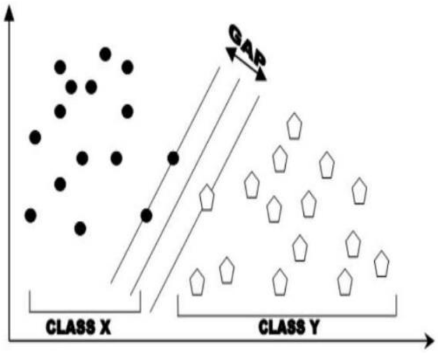
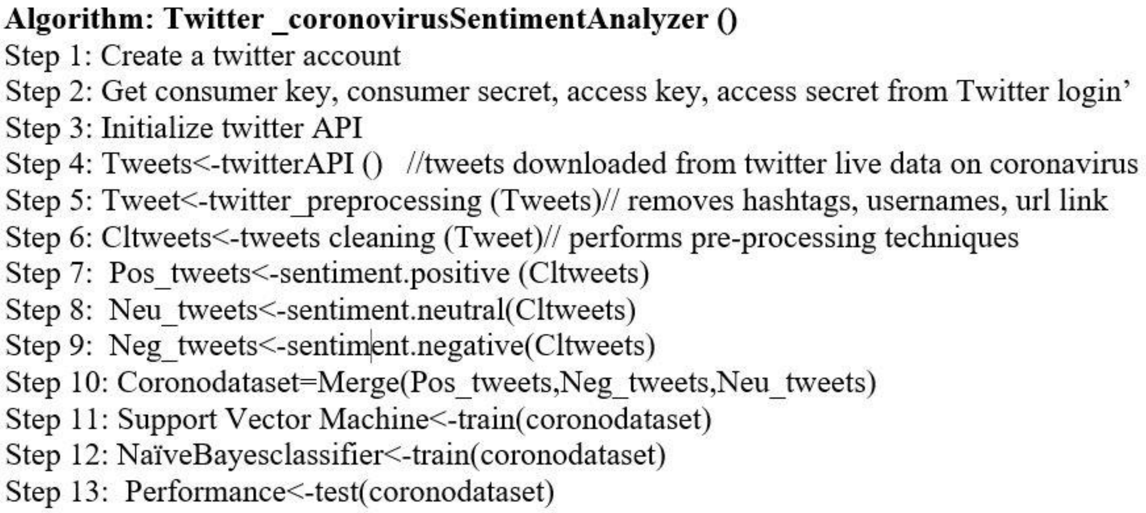


FIGURE 3. Representation of Support Vector Machine.

Lexicon-Based Model

In order to draw conclusions about feelings, this method analyses a word or group of words. Such methods are referred to as keyword-based methods. Dictionary-based and corpus-based lexicon usage are the two categories. It includes a fast categorization rate.

This method makes assumptions the feelings based on a word or group of words. These methods are referred to as keyword-based methods. Dictionary-based and corpus lexicon systems are both used. If the polarization is “positive, negative, or impartial” it can quickly analyse the data using this approach. [7] Fundamentally said, lexicon groups can be either dictionary- or corpus-based. [8] [9]. The dictionary approach uses an existing vocabulary, whereas corpus-based analysis considers the likelihood that a sentiment term would appear alongside a positive or negative group of words [10].



```
Algorithm: Twitter _coronovirusSentimentAnalyzer ()
Step 1: Create a twitter account
Step 2: Get consumer key, consumer secret, access key, access secret from Twitter login'
Step 3: Initialize twitter API
Step 4: Tweets<-twitterAPI ()  //tweets downloaded from twitter live data on coronavirus
Step 5: Tweet<-twitter_preprocessing (Tweets)// removes hashtags, usernames, url link
Step 6: Cltweets<-tweets cleaning (Tweet)// performs pre-processing techniques
Step 7:  Pos_tweets<-sentiment.positive (Cltweets)
Step 8:  Neu_tweets<-sentiment.neutral(Cltweets)
Step 9:  Neg_tweets<-sentiment.negative(Cltweets)
Step 10: Coronodataset=Merge(Pos_tweets,Neg_tweets,Neu_tweets)
Step 11: Support Vector Machine<-train(coronodataset)
Step 12: NaïveBayesclassifier<-train(coronodataset)
Step 13:  Performance<-test(coronodataset)
```

Algorithm 2 : Twitter _CoronoSentimentAnalyzer

From the above algorithm(2) the live tweets are obtained from Twitter using the twitter API for the mentioned algorithms. The functions tweet preprocessing are used to remove the hastags, usernames, and urls. The tweets cleaning( ) function tokenizes and stems the output of the function. Using the methods sentiment postive( ), sentiment negative (), and sentiment neutral(), the tweets are divided into positive, negative, and neutral categories( ). Using the Merge() method, the tagged tweets are combined into the Corono dataset.

RESULT AND DISCUSSION

To implement the machine learning classifier requires Python. By developing a Twitter application that communicates with the Twitter API and downloads tweets using phrases that match the search terms, twitter data is gathered. Confidential keys such a consumer key, consumer secret, access token, and access token secret are created using the twitter API. These keys aid in user authentication for Twitter API access. Read-only access is the default setting for Twitter accounts. With the use of these keys, certain key phrases from Twitter, such "Corono Virus," are filtered in the appropriate locales and languages.

By combining sentiment analysis and machine learning to study the gathered tweets during the outcomes of the corono virus outbreak.In the experimental setup, the performance accuracy was tested using the metrics like Precision, Recall, and F1-score using both supervised and unsupervised lexicon models. From the above Table (1) (2) (3) the metrics like Precision, F1-Score, Recall, and, Accuracy were calculated. It can be seen that the lexicon model metrics like Textblob and Vader are nearly similar in accuracy with 85% and 84%.

TABLE 2. Performance Evaluation of Lexicon Models

COMPARISON OF LEXICON MODELS				
LEXICON MODEL	**PRECISION**	**RECALL**	**F1-SCORE**	**ACCURACY**
TEXTBLOB	82%	85%	89%	85%
VADER	89%	80%	82%	84%

From Figure 4. the accuracy analysis of Lexicon models was represented clearly and compared with graphical representation.

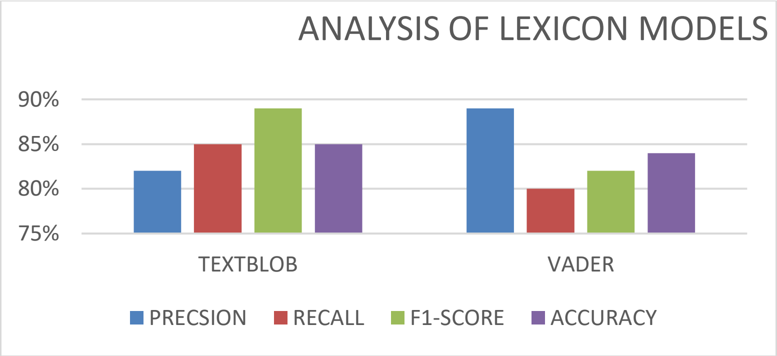


FIGURE 4. Accuracy Analysis of Lexicon Models

In Table 2 and Table 3, “Naive Bayes and Support Vector Machine” has related as well as the performance metrics. Utilizing feature extraction technique of Bag of Words (BOW), Naive Bayes outperforms SVM with an accuracy of 86%. From Fig (5) the accuracy analysis using Bow was represented graphically by means of a bar chart.

TABLE 3. Performance Evaluation of Machine learning Classifier Using BOW

ALGORITHM	PRECISION	RECALL	F1-SCORE	ACCURACY
NB	86%	82%	84%	86%
SVM	88%	82%	87%	83%

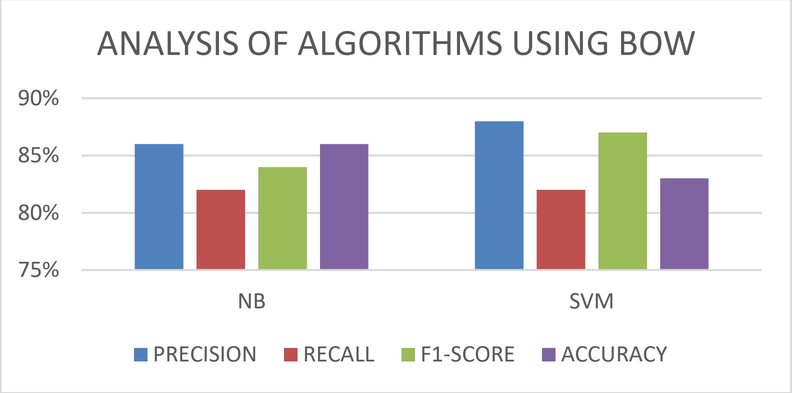


FIGURE 5. Accuracy Analysis using BOW

In table 3, the TF-IDF was applied using the NB and SVM classifiers. With an accuracy of 88%, SVM outperforms NB. From Fig (6) the accuracy analysis using TF-IDF was represented graphically.

TABLE 4. Performance Evaluation of Machine Learning Classifier using TF-IDF

ALGORITHM	PRECISION	RECALL	F1-SCORE	ACCURACY
NB	88%	86%	84%	87%
SVM	90%	84%	86%	88%

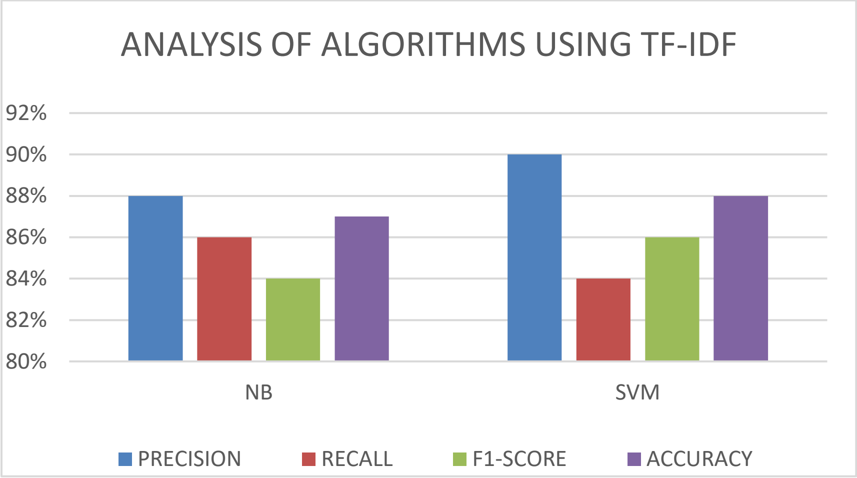


FIGURE 6. Accuracy Analysis of TF-IDF

Word Cloud Representation of COVID Data.

In the Word Cloud below can see the keywords that appeared most frequently in tweets related to COVID-19.

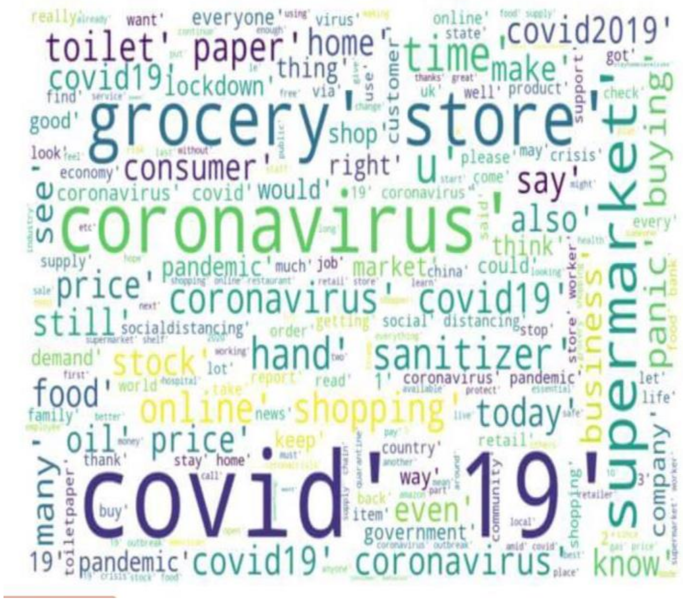


FIGURE 2. Word Cloud Representation of Covid data

Word cloud also called a tag cloud or text cloud, the larger an individual word is, the more often it is mentioned. Among environmental professionals, students, and researchers, the words "life", "Covid", "Lock down", "job" ,” Sanitizer” and "things" appear most frequently.

CONCLUSION

In this study, both supervised and unsupervised models are used to analyze sentiment. The SVM model using TF-IDF features, a machine classification approach is the best with 88% accuracy. According to the results, the Machine Learning model is more perfect than both Text blob and VADER lexicon. By comparing the top models from both strategies, related to lexical models, predictable supervised models perform better. In the future, more characteristics should be extracted using “machine learning classifiers like Random forest, Logistic regression, and moreover deep learning”, which can also be utilized to increase the accuracy level.

REFERENCES

[1]. H Hassan Raza M Faizan, A Hamza, A Mushtaq, N Akhtar, "Scientific Text Sentiment analysis using Machine Learning Techniques", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol.10, No.12, (2019). https://doi.org/10.14569/IJACSA.2019.0101222

[2]. Nitesh Sharma, Ussama Yaqub, Vijayalakshmi Atluri, Soon Ae Chun, Rachit Pabreja, and Jaideep Vaidya "Web-based Application for Sentiment Analysis of live tweets" ACM, https://doi.org/10.1145/3209281.3209402

[3]. Oumayama EI Ansari, Hajar Mousannif and Jihand Zahir "Context-Based Sentiment Analysis: A Survey" https://doi.org/10.1007/978-3-030-02852-7_8

[4]. B Das, S Chakraborty, "An Improved Text Sentiment Classification Model Using TF-IDF and Next Word /Negation", (2018)

[5]. Ashwin Sanjay Neogi, Kriti Anilkumar Garg, Ram Krishn Mishra, Yogesh K Dwivedi, "Sentiment Analysis and classification of Indian Farmers protest using Twitter, Elsevier,2021,pp:1-10. https://doi.org/10.1016/j.jjimei.2021.100019

[6]. Abdul Mohaimin Rahat, Abdul Kahir, Abu Kaisar Masum, "Comparison of Naïve Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset, IEEE,(2019), pp: 266-270.

[7]. T. Nikil Prakash and A.Aloysius, Applications, Approaches and Challenges in Sentiment Analysis(AACSA), IRJMETS 02(07)(2020),pp:910-915.

[8]. R.Aishwarya, A. AShwatha, C.Deepthi and Beschi Raja, A Novel Adaptable Approach for Sentiment Analysis, IJSCSEIT(2019),pp254-263. https://doi.org/10.32628/CSEIT195263

[9]. Vishal A.Kharde and S.S.Sonawane, Sentiment Analysis of Twitter data: A survey of techniques, International Journal of Computer Applications 139(11)(2016),1-10. https://doi.org/10.5120/ijca2016908625

[10]. R. Cynthia Monica Priya and Dr.J.G.R. Sathiaseelan, An explorative study on sentiment analysis,19th Oct (2017), WCCCT,140-142.

[11]. Ankur Goel, Jyoti Gautam and Sitesh Kumar, Real-Time sentiment analysis of tweets using Naïve Bayes,2nd International Conference on NGCT Dehradun, India(2016),pp-257-261.

[12]. Monika Kabir, Mir MD Jahangir Kabir, Xu Shuxiang and Bodrunnessa Badhon, An empirical research on sentiment analysis using machine learning approaches, International Journey of Computer Applications,(2019). https://doi.org/10.1080/1206212X.2019.1643584

[13]. Ghulam Musa Raza, Zainab Saeed Butt, Seemab Latif, Abdul Wahid,"Sentiment Analysis on COVID Tweets:An Experimental Analysis on the Impact of Count Vectorizer and TF-IDF on Sentiment Predictions using Deep Learning Models",IEEE,(2021)

[14]. Mayur Wankhade, Annavarapu Chandra Sekhara Rao, Chaitanya Kulkarni," A Survey on Sentiment Analysis methods, Applications and Challenges", Artificial Intelligence Review, Springer,(2022)pp:5732-5780. https://doi.org/10.1007/s10462-022-10144-1

[15]. Miftahul Qorib, Timothy Oladunni, Max Denis, Esther Ososanya, Paul Cotae," Covid -19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter data set," Expert Systems with Applications, Elsevier,(2023).

[16]. D. Rajeswaro Rao, S.Usha, S.Sri Krishna, M.Sai Ramya, G.Sri Charan, and U.Jeevan, Result prediction for political parties Twitter sentiment analysis, International Journal of Computer Engineering and Technology(IJCET) 11(4),(2020),pp:1-6.

[17]. Rupinder Kaur, Rajvir Kaur, Manpreet Singh and Dr.Sndeep Ranjan, Twitter sentiment analysis of the Indian union Budget, International Journal of Advanced Science and Technology,(2020),pp:2282-2288.

[18]. Reena G.Bhati, A survey on sentiment analysis algorithms and datasets, Review of Computer Engineering Research,(2019),pp-84-91 https://doi.org/10.18488/journal.76.2019.62.84.91

[19]. T.Nikil Prakash, A.Aloysius," Lexicon Based Sentiment Analysis(LBSA) to Improve the Accuracy of Acronyms, Emotions, and Contextual Words", Statistics and Application,(2022),pp-75-87.

[20]. A.Sathya, M.S Mythili,"An Investigation of Machine Learning Algorithm in sentiment Analysis,"Advances and Applications in Mathematical Sciences,2022,pp.4575-4584.

[21]. Hassan Adamu,Syaheerah Lebai Lutfi,Nural Hashimah Ahamed Hassain Malim,Rohail Hassan,AssuntaDi Vaio and Ahmad Sufril Azllan Mohamed,Framing twitter public sentiment on Nigerian government COVID-19 palliatives distribution using machine learning,Sustainability 13,no.6,2021. https://doi.org/10.3390/su13063497

[22]. Nalini Chitalapudi,Gopi Battineni and Francesco Amenta,"Sentimental Analysis of COVID -19 Tweets Using Deep Learning Models",Infectious Disease Reports, 2021,pp.329-339. https://doi.org/10.3390/idr13020032

[23]. O. Baker, J. Liu, M. Gosai and S. Sitoula, "Twitter Sentiment Analysis using Machine Learning Algorithms for COVID-19 Outbreak in New Zealand," 2021 IEEE 11th International Conference on System Engineering and Technology (ICSET), 2021, pp.286-291 https://doi.org/10.1109/ICSET53708.2021.9612431

[24]. Ishaani priyadarshini,pinaki Mohanty,Raghvenda kumar,Rohit Sharma,Vikram puri,Pradeep kumar singh, "Multimedia Tools and Applications,Springer,2022,pp.27009-27031. https://doi.org/10.1007/s11042-021-11004-w