Journal of Autonomous Intelligence (2024) Volume 7 Issue 2
doi: 10.32629/jai.v7i2.801

ORIGINAL RESEARCH ARTICLE

Sentiment analysis and classification of COVID-19 tweets using machine learning classifier

Chataparti Suvarna Lakshmi[1], Sameer Saxena[1], Billakurthi Suresh Kumar[2,*]

[1] *Amity Institute of Information Technology, Amity University Rajasthan, Jaipur, 303002, India*

[2] *School of Computer Science and Engineering, Sanjay Ghodawat University, Kolhapur, Maharashtra, 416118, India*

*** Corresponding author:** Billakurthi Suresh Kumar, sureshkumarbillakurthi@gmail.com

ABSTRACT

In March of 2020, the World Health Organization identified COVID-19 as a new pandemic and issued a statement to that effect. This fatal virus was able to disperse and propagate throughout several countries all over the world. During the progression of the pandemic, social networking sites like Twitter generated significant and substantial volumes of data that helped improve the quality of decisions pertaining to health care applications. In this paper, we proposed a sentiment classification using various feature extraction and machine leavening techniques for social media dataset. The system has divided into four phase data collection, preprocessing and normalization, feature extraction and feature selection and finally classification. In first phase we collect data from social media sources such as twitter using Twitter API. In second phase the tweets, data was ready for preprocessing and it was sorted into three categories: positive, neutral, and negative. During the third phase, various features were extracted from the tweets by employing a number of widely utilized approaches, including as bag of words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, and FastText, to gather feature datasets. These methods were employed to extract distinct datasets for the features. The final phase different machine learning classification algorithms are applied for detection of sentiment using machine learning. In the extensive experimental analysis, the BoW performed better results with modified support vector machine (mSVM) than existing machine learning algorithms. The proposed mSVM performed superiorly to the other classifiers by 98.15% accuracy rate. Once the tweets are correctly classified as COVID-19 tweets, it is further categorized into three sentiments that is positive, negative and neural. Proposed mSVM achieves 93% of accuracy rate for positive sentiment which better as compared to other Machine Learning (ML) classifiers.

Keywords: sentiment analysis; COVID-19 tweets; machine learning; modified support vector machine

ARTICLE INFO

Received: 14 July 2023
Accepted: 5 September 2023
Available online: 12 December 2023

1. Introduction

Since the beginning of 2020, COVID-19 has spread quickly over the world, having an impact in every region. The WHO declare the COVID-19 is the pandemic disease including social media websites over the world. The pandemic disease is a type of disease that spreads across a large geographic area, affecting a significant portion of the population and often spanning multiple countries or even continents. Pandemic diseases are characterized by their ability to cause widespread illness, social disruption, and sometimes significant mortality.

The COVID various was passed into one person to another person very rapidly[1]. This strange disease, whose existence has been recorded in millions of cases, has been responsible for the deaths of countless of people. This was awful news for all, especially for authorities worldwide, and it quickly spread around the world. The identification of public sentiment was important for COVID-19 vaccine during that period[2]. It was a challenge for all nations to determine how to prevent the disease from spreading as well as how rescue their nation, but the countries that were hit the hardest by the pandemic decided to put their nations under an entire lockdown[3]. A sudden decision was made by all nations to implement either a complete or partial lockdown, and it is believed that approximately 10 million individuals became infected[4].

This quarantine may have been successful in slowing the transmission of the COVID-19 virus, however it has also produced a multitude of other difficulties, such as unemployment, starvation, hyperinflation, and other social concerns[5]. The various machine learning algorithm and hybrid feature extraction techniques are applied by existing researchers for detection of human emotions during the pandemic. The NLP techniques are used for extraction of features and supervised machine learning classification algorithms are used for detection of sentiment. Some studies indicate that the new virus and lockdown have led to a significant increase in the number of individuals seeking jobs, with an approximated 10 million jobless people all across the world[6]. Since everyone was dealing with a variety of personal and societal problems at the same time, the lockdown had an influence on the human brain. Aside from this, many people appreciated the chance to spend quality time with their family during the lockdown and were pleased with the decision made by the administration.

The various approaches are proposed for sentiment classification on COVID-19 using various Natural Language Processing (NLP) and machine learning methods[7–10]. The similar deep learning approaches also describes for sentiment detection on COVID-19 vaccine[11,12]. In light of the circumstances, it was extremely challenging to make decisions regarding how to solve the problems. The various machine learning algorithms and feature selection techniques are applied in Baker et al.[13] and similar approach has proposed for geographical revision based sentiment classification[14]. The hybrid feature extraction and selection techniques are proposed by Andhale et al.[15] for classification on Twitter data. They have used unstructured text data for identification of sentiment users.

The major focus to evaluate regional data from twitter and other social media application using various feature extraction, optimization and classification techniques are proposed[16–17]. Regardless of the reality that COVID-19 is a subject that has only been around for a short while, several contributions have already been made to its many different components[18,19]. In order to identify the sentiments of end users and emotions hybrid sentiment classification techniques are proposed including image and text dataset. COVID-19 has decided to embark on an extensive research project using deep learning (DL), artificial intelligence (AI), and machine learning (ML)[20]. Even though there has been a lot of research done on the COVID-19 immunization, the health risks are still the same[21,22].The major objective of this research to extract the social media dataset for identification of sentiment. Various feature extraction techniques are carried out such as TF-IDF, BoW, lemmas, relation dependencies, etc. are used for module training. Finally various machine learning algorithms such as SVM, Naïve Bayes, random forest, J48, ADABOOST, ANN are applied for detection for sentiment. The emotions conveyed by the tweets were classified as positive, negative, or neutral in order to facilitate the feature extraction process. The data from tweets about COVID-19 that were put out when the pandemic was going on were analysed by the researchers so that they could look at trends and impacts. Because Twitter is widely regarded as among the most credible social media channels, it has been utilised in this process to categorise and organise the thoughts. One more thing to consider is that many people stayed at home during the lockdown and stayed connected to the internet the whole time.

Even though the dataset contains tweets that are linked to COVID-19, it nevertheless reflects the perspectives of individuals from a variety of various contexts and from a wide range of different backgrounds. The dataset on Twitter was obtained via Kaggle, and it contains around 170,118 tweets that were posted at some point during the outbreak from users located all over the world. Conventional machine learning classifiers were utilised for the training of 70 percent of the tweets. On the other hand, evaluation criteria such as accuracy, recall, precision, and F1-score were used to examine 30 percent of the data for problems associated with categorization. The following is an overview of the significant contributions that the proposed effort will make:

- Utilizing COVID-19 to investigate the feelings of people while taking into account neutral, positive, and negative tags.
- Presenting and comparing the 4 distinct feature extraction techniques namely bag of words, TF-IDF, Word2Vec and FastText using conventional machine learning classifiers and a proposed mSVM classification model.
- Finding out how COVID-19 affects humans, diagnosing and treating anxiety and panic attacks, and preventing future outbreaks.
- Discovering the general public's perspective on the epidemic in order to facilitate the decision-making process and make it much simpler.

The remaining portions of this article are structured as follows: section 2 shows some related work of sentiment analysis of COVID-19 posts on social networking sites using a variety of machine or deep learning classifiers. The architecture of proposed mSVM model for recognising and analysing COVID-19 tweets is presented in section 3. In part 4, the experimental results and discussions are illustrated and in section 5, the conclusion and the future scope of this research is discussed.

2. Literature survey

Natural language processing (NLP) is the field of study that encompasses the technology that is used for sentiment analysis, as stated by Zope and Rajeswari[1]. Analysis of sentiment is utilised in both behavior prediction and evaluation. For the purpose of conducting sentiment analysis using the dataset of COVID-19 tweets, the machine learning techniques of random forest classifier, decision tree and support vector machine are all under investigation. There was a total of 179,108 tweets incorporated into this study's findings. The first step in cleaning and analysing these tweets is called pre-processing. The specific perceptions of those who were impacted by the epidemic are revealed through the sentiment analysis of COVID-19 twitter posts that were used in this study. The major goal of this research is to identify the sentiment of social media users using machine learning techniques. It is beneficial to have a deeper understanding of the feelings of people, particularly during the time of an epidemic. Twitter, which is a platform for microblogging, has a sizeable set of datasets that represent a broad spectrum of human emotions. These datasets include examples of fear, joyful, sad, angry, and joyous expressions, amongst others. The importance of sentiment analysis cannot be overstated when attempting to ascertain the general consensus of the population. The findings of this research can assist in determining how a pandemic might influence the choices that members of the general population make.

Through the application of machine learning, Chitra et al.[2] investigates people's perspectives on the COVID-19 vaccine amongst university students. An investigation employing social network analysis was conducted about the COVID-19 vaccination. The ML algorithms such as Naïve Bayes, SVM and logistic regression are three methods that can be used to conduct sentiment analysis. These methods offer the greatest potential for accurate conclusions.

Over 18 million tweets linked to the coronavirus were collected and analysed by Soomro et al.[3] for their study. The tweets were collected during 3 months from March 2020. In order to determine the nature of the connection that exists between public opinion and the total number of COVID-19 instances, a sentiment analysis was carried out with the use of a rule-based supervised machine learning technique known as VADER. It also takes into account the amount of times a country is mentioned in tweets as well as the rise in the daily amount of COVID-19 instances in that country. One of the conclusions is that there is an association between the number of twitter posts that mention Italy, the United States of America, and the United Kingdom and the daily growth in new COVID-19 cases in those countries. Other findings include this correlation.

Adamu et al.[4] mentioned the utilization of text analytics and sentiment analysis in NLP based on tweets in examine public sentiment as well as derive insights concerning COVID-19 vaccines in the medical sector. These methodologies were applied to the analysis of tweets. In order to identify and analyze the data, we used two different machine learning algorithms: SVM and k-nearest neighbor. In order to assist in the detection of the public mood utilizing the three sentiment polarity groups—favorable, unfavorable, and neutral—a variety of pre-processing approaches were implemented. The outcome of the distribution of sentiment classes suggests that 31% of the public mood for COVID-19 vaccinations is good, 22% of the public mood is negative, and the rest 47% were categorized as neutral emotion. The results of the experiments using machine learning algorithms show that SVM achieved an accuracy of 88%, which is higher than the accuracy achieved by KNN (78%).

Tareq and Hewahi[5] analyzed and assessed the emotion of the tweets that were shared during the COVID-19 pandemic, as well as experimenting the bidirectional long term short memory (BLTSM) of recurrent neural network in forecasting the emotion class, which are negative, positive, and unbiased, respectively. The findings indicate that there is only a tiny difference between the favorable and negative tweets, which highlights the necessity for attention to be paid in order to boost awareness and, as a result, decrease negativity. In addition, the BLSTM was able to accurately predict the sentiment categories (negative, positive, and neutral), with an accuracy rate of 86.15%. The great accuracy leads one to the conclusion that BLSTM is capable of adapting to new situations and predicting the emotion of a text.

Tao et al.[6] performs COVID-19 related MLSA on a variety of internet media sources in order to establish the relative usefulness of each for extracting public sentiment. A neural network with long short-term memory (LSTM) is utilised in order to carry out the natural language processing. When the trends of sentiment on Twitter, Reddit, and USA Today are compared to those of a control survey conducted by the data intelligence corporation Daytime Consult, it is discovered that Twitter has the smallest deviation in patterns compared to that of the regulate survey. Other social media platforms, such as Reddit, are also included in the comparison. Assuming that the control is objective, Twitter is a better indication of popular opinion in comparison to Reddit and USA Today. This makes Twitter suitable for potential future uses of MLSA, particularly when used in conjunction with surveys that have already been conducted. This paper contributes to the advancement of research in MLSA, which has consequences for making educated judgements regarding COVID-19 treatment and recovery, smoothing future potential epidemic curves, and signalling trends in general psychological and mental health.

Khan et al.[7] demonstrate how TF-IDF features can improve the overall effectiveness of supervised machine learning techniques. Furthermore, in the work that was presented, the gradient boosting machine beats all of them and obtains a high accuracy of 96% when combined with TF-IDF features. The purpose of this investigation was to investigate how people living in the United States are responding to the current predicament. The successful completion of the experiments provides evidence that the strategy is effective.

The purpose of the sentiment analysis that Jannah and Hermawan[8] proposed an approach sentiment classification using machine leavening on YouTube comments. The data that was used was extracted from YouTube comments by searching for the term ‘COVID-19 vaccine’ on the YouTube channels of the Indonesian Department of Health, the Presidency Secretariat, President Joko Widodo, as well as Najwa Shihab. The time period for retrieving feedback data is from November 2020 to July 2021. Python is the language of programming that is used in conjunction with the access token that was obtained from the

YouTube API in order to perform the manual crawl of the data in order to gain the data that was requested. The comment data is broken down into three different groupings: neutral, positive, and negative. Applying a number of different machine learning techniques, including k-nearest neighbor, SVM, DT, RF, AdaBoost, Multinomial NaïveBayes, logistic regression and stochastic gradient descent, will help discover the optimum classification model. With sentiment metrics obtained from more than 4.6 thousand negative responses (44.9%), more than 3.4 thousand good responses (33.3%), and the remaining more than 2.2 thousand impartial comments (21.8%), the results of sentiment classification research pertaining to COVID-19 tend to get bad comments. A value of 68% was found to be the optimal level of accuracy when using the linear regression technique.

Using tweets as a research tool, Patravali and Algur[9] proposed a system identify the sentiment and emotions during the pandemic using machine learning techniques. Twitter posts sent out by users in India in 2020 were compiled throughout the following time periods: April, June and August 2020, in order to investigate shifts in sentiment and emotion. Techniques of sentiment analysis from TextBlob and NRC lexicon are compared in the work that is proposed. TextBlob does a better job at classifying sentiments, while the NRC Lexicon method analyses emotional states in greater detail. The work that is being suggested investigates the differences in sentiment that occur over three distinct time windows for tweets and examines the findings obtained from both experimental techniques. The results obtained from both of the approaches reveal that favorable sentiments predominate over negative sentiments, while neutral sentiments make up the bulk of the sample.

Sancoko et al.[10] intended to develop an ensemble learning method that is capable of classifying the sentiment contained within the people's opinions obtained from Twitter. The Naïve Bayes (NB), C4.5, and k-nearest neighbors algorithms were utilised as the basic learners in the ensemble classifier, and a voting method was used to create the final conclusion. The ensemble model made use of a dataset of 3884 data samples that was effectively obtained using Twitter API and processed using the TF-IDF method with the objective of learning. These data were relevant to the prevention of the COVID-19 outbreak. The mood of the view expressed in each piece of data is represented in the dataset by one of two classes: either "positive" or "negative." After being examined with 10-fold cross validation, the suggested framework received a score of 81.20 percent for its precision, 79.49 percent for its recall, and 80.61% for its accuracy. When contrasted to the several learning techniques that just used a single machine-learning algorithm, it performed far better.

Balaji et al.[11] gathered the general population of India's thoughts on vaccines by reading their tweets on Twitter. The most cutting-edge machine learning and deep learning techniques were applied to the analysis of over four hundred thousand tweets relating to vaccines that were posted between 4 May and 11 May 2021, as well as between 13 August and 21 August 2021. When compared to the results obtained by other models on the gathered Twitter dataset, the BERT and RoBERTa models generated some encouraging findings.

Using a method based on machine learning, Islam et al.[12] provide an estimation of the effect that the COVID-19 outbreak had on the mood of the people in Bangladesh. In order to accomplish this objective, COVID-19 tweets were gathered over the course of a predetermined time period, and a deep learning classifier with an area under the curve of 0.76 was then constructed. The study examines the transmission of the disease and calculates the various emotions felt by the general people during the outbreak. And demonstrates that a sizeable number of respondents, 55%, had a bearish sentiment towards COVID-19, whereas 38% and 7% of people, correspondingly, had a good sentiment or were indifferent. According to the findings of this study, there is a correlation between the number of active COVID-19 cases and the degree to which people participate in social media. In addition, the findings of this study revealed people's attitudes toward a number of significant worries in relation to the COVID-19 pandemic.

The research conducted by Baker et al.[13] uses data from Twitter to conduct a sentiment classification of the COVID-19 pandemic spread in New Zealand. The studies were produced using a number of different categorization methods using machine learning, in particular NB, k-nearest neighbor, convolutional neural network, and SVM. The researchers also put these algorithms through their paces in Python and RapidMiner, two distinct data mining platforms. After doing so, the metrics was compared that were obtained from the various techniques and platforms in order to determine which algorithm was best suited for sentiment classification. The researchers conclude by providing an illustration of the experimental findings that demonstrate the performance of NaïveBayes and SVM. These results reveal a longer processing time but resulted to an enhanced Twitter sentiment evaluation results that outperformed the results obtained by the other algorithms. After that, verify the usefulness of these models by contrasting the outcomes of running the identical models in Python and RapidMiner.

Nwafor et al.[14] present a sentiment classification of Twitter conversations around the use of the COVID-19 vaccination. It concentrates on key locations of the United States, notably urban areas with significant numbers of people of African American descent. As a means of generating baseline models, a number of different machines learning techniques, including logistic regression, SVM, NB were put to use. In addition, very accurate Transformer-based sentiment categorization models have been built. These models were fine-tuned using natural language processing techniques. The findings of the investigation indicate that fine-tuning the dataset using a transformer-based paradigm, COVID-BERT v2, outperforms than the baseline methods; although, the accuracy is still very low despite this improvement. One possible explanation for this is because the dataset used for training is quite little.

Andhale et al.[15] intend to implement data analysis and innovative machine learning approaches such as robustly optimised BERT pre-training approach (RoBERTa) and CNN-RoBERTa sentiment retrieval to integrate methods such as sentiment classification, frequency distribution and comparative evaluation on information recorded from social networking sites for effective comparison on the consequences of COVID-19 in India while also constructing a web-based interface for efficacious exposition and deep insight over the crisis.

During the COVID-19 outbreak, Mohsen et al.[16] advocated using machine learning (ML) techniques for Arabic Sentiment Classification to identify the positive and negative ideas connected to isolation and social distance. This was done in order to gain a better understanding of the opinions. The various essential and ensemble machine learning classifiers were supplied, and their efficacy in categorizing the collected unbalanced dataset was compared. In addition, a number of different synthetic minority over-sampling techniques (SMOTe) were used to the class imbalanced in order to conduct an analysis of their effectiveness. According to the findings, the SMOTe edited nearest neighbor (SMOTEENN) strategy performed significantly better than the other SMOTe methods. In addition, the findings demonstrated that the use of ensemble classifiers is superior to the use of single classifiers when dealing with imbalanced datasets. When using SMOTEENN, the average rating of the F1 score of classification models is more reliable than the F1 score of ensemble classification models.

3. Research methodology

The architecture of proposed mSVM for detecting and analysing the COVID-19 tweets is depicted in **Figure 1**. The proposed framework is divided into several stages. In the first stage, data from Twitter is retrieved through Kaggle, which is accessible to the public. In the second stage, the data will go through a pre-processing process in order to be cleansed and prepared for analysis. Examples of preprocessing process include removing stopwords, removing numerical values, removing punctuation, stemming, etc. During the third stage, the data is divided into a ratio of 70:30 for training and testing. The training dataset contains 70 percent of the total data, whereas the testing dataset contains 30 percent. The various features of the dataset are extracted by making use of four different techniques for feature extraction. These methodologies are referred to as bag of words, TF-IDF, Word2Vec, and FastText In the fourth stage, data is classified by using traditional ML classifiers in addition to the proposed mSVM. These classified data are then subjected to further analysis with the help of performance metrics such as recall, accuracy, f-score and precision. The detail description of each stage of proposed model is explained as follows.

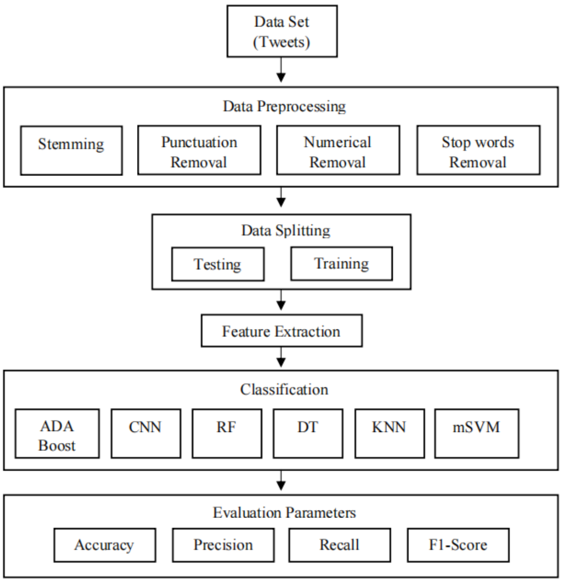

Figure 1. Framework of proposed mSVM classifier for sentiment analysis and classification of COVID-19 tweets.

Data collection: In any statistical investigation, one of the most challenging tasks is dataset preparation. It is generally agreed that this stage of the data analytics life cycle requires the maximum amount of time to complete. The gathering of as much information as is feasible is the initial step that must be taken in this phase. For the purpose of this study, data taken from Twitter was gathered via Kaggle.

Data pre-processing: In order to use the data that was collected, the raw data that was obtained must first be transformed into a form that is suitable for use with machine learning techniques. This process is called as data pre-processing and cleaning. While preprocessing and cleaning data, one of the tasks that must be completed is an examination of the data to determine whether or not it contains any errors that would render it unfit for use in later processes. The data may have problems such as duplication, erroneous data, values that are missing, data that is formatted incorrectly, and so on. In most cases, the data gathered from social media platforms is not yet ready for analysis. The data needs to go through a series of preprocessing procedures in order to be cleaned up and made suitable for analysis. These steps include deleting data that is not linked to COVID-19, eliminating duplicates, and eliminating punctuation, etc. The elimination of stopwords, punctuation, and numerical values, as well as stemming are all typical examples of preprocessing procedures.

Before using machine learning algorithms to the collected data in order to derive insights from them, it is necessary to generate a training dataset of a high quality. The labels for this dataset need to be entered by hand by a human. The effectiveness of the machine learning algorithms is directly proportional to the quality of the training dataset that has been created. Every tweet has been given a classification indicating whether it is positive, negative, or neutral in terms of the sentiment it expresses. The final dataset contained a total of 6903 tweets that had been properly classified after being cleaned, pre-processed, and structured. The labelled dataset that was produced as a consequence was used as input for machine learning techniques, which will be able to automatically determine the sentiment of tweets that will be collected in the future.

Feature extraction: This phase intends to make use of machine learning algorithms so that it is possible to construct a model that can predict the sentiment of tweets that are relevant to COVID-19. Nevertheless, the algorithms that make up machine learning can't directly cope with textual data. As a consequence of this, the process of extracting features is an essential stage that must be completed before developing the machine learning models. In the beginning, fundamental numerical features were retrieved, such as the total number of words, the length of tweets, the average word length, the amount of hashtags, and so on. In addition, advanced methods were utilised in order to extract features, which are explained as follows.

Traditional bag-of-words (BoW): The BoW model is necessary for encoding data in natural language and retrieving that data. A text is seen through the perspective of this paradigm as merely a collection of its terms. Grammar and word order are disregarded, but multiplicity is maintained.

Term frequency-inverse document frequency (TF-IDF): A scoring measure that is utilized in information retrieval is known as the TF-IDF (IR). The major purpose of using TF-IDF is to draw attention to the importance of a certain word inside a specific text. Incorporating the following metrics allows for the computation of the definition: (i) the number of times a word occurs in a text; and (ii) the word's inverse document frequency across a set of texts.

Word2Vec: The method of learning word embedding by making use of neural networks which is utilised the most frequently is called Word2Vec. The trained model carried out a mathematical calculation on the text corpus in order to group together terms that were mathematically equivalent. Both the skip-gram-based technique and the continuous bag of words (CBOW) technique are important components of Word2Vec. The skip-gram-based technique involves making predictions about the context based on a single word. The CBOW method makes predictions about the term according to the context. This research led to the development of CBOW, which was then used to train on the corpus with the following parameters: dimension = 100, least word frequency = 5, and window lengths = 5.

FastText: The Facebook team offers a service known as FastText, which is a method of word embedding that is focused on the skip-gram model and in which every word is converted into an N-grams character. Even misspelt or uncommon words that are not included in the dictionary will have an embedding since the terms in the training data are linked to a vector representation that is the summation of every character's N-gram. When applied, the pre-trained FastText embedding techniques result in the production of one vector for every word contained within a particular tweet. When compared to Word2Vec, the accuracy of FastText, which is an upgrade to that programme, has been found to be higher.

Classification and evaluation: In the stage of classification, several ML classifiers are used. Those classifiers are k-nearest neighbor, random forest, Adaboost, convolutional neural network, and decision tree. These classified data are then subjected to further analysis with the help of performance metrics like as recall, accuracy, F-Measure, and precision

4. Result and discussion

The traditional ML classifiers and proposed mSVM model is trained using unique feature extraction techniques namely bag of words, TF-IDF, FastText and Word2Vec. The performance analysis of the classification models is performed using these feature extraction methods which is depicted in the following **Table 1**.

It is observed from the above table that the performance of proposed mSVM model using FastText extraction technique is better as compared to conventional ML classifiers.

Table 1. Performance analysis of traditional ML classifiers and proposed mSVM using feature extraction methods.

Feature extraction	**Classifiers**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Bag of words	Adaboost	92.54%	94.03%	88.14%	90.73%
	CNN	92.87%	94.26%	88.52%	91.19%
	RF	89.39%	88.18%	85.67%	86.28%
	DT	87.36%	92.15%	80.19%	84.92%
	KNN	94.58%	93.65%	93.12%	92.36%
	mSVM (proposed)	**98.15%**	**96.88%**	**96.35%**	**97.53%**
TF-IDF	ADABOOST	89.69%	92.31%	83.12%	86.78%
	CNN	92.61%	94.05%	88.11%	90.56%
	RF	89.73%	90.53%	83.48%	86.44%
	DT	89.21%	92.01%	82.79%	86.59%
	KNN	94.52%	94.35%	93.16%	93.68%
	mSVM (proposed)	**96.10%**	**96.79%**	**94.09%**	**95.37%**
FastText	ADABOOST	91.36%	92.18%	87.07%	89.66%
	CNN	91.58%	93.33%	87.43%	90.14%
	RF	88.29%	87.12%	84.56%	85.16%
	DT	86.18%	91.11%	79.14%	83.78%
	KNN	93.34%	92.23%	92.31%	91.12%
	mSVM (proposed)	**97.11%**	**95.56%**	**95.13%**	**96.46%**
Word2Vec	ADABOOST	88.52%	91.17%	82.07%	85.64%
	CNN	91.56%	93.02%	87.09%	89.48%
	RF	88.64%	89.48%	82.37%	85.39%
	DT	88.14%	91.26%	81.68%	85.84%
	KNN	93.43%	93.22%	92.13%	92.59%
	mSVM (proposed)	**95.23%**	**95.91%**	**93.32%**	**95.37%**

The **Figure 2**, depicts the analysis of accuracy of proposed mSVM and traditional ML classifiers using various feature extraction methods. It is observed that the accuracy of mSVM using bag of words feature extraction method is 98.15% which is better as compared to other ML classifiers.

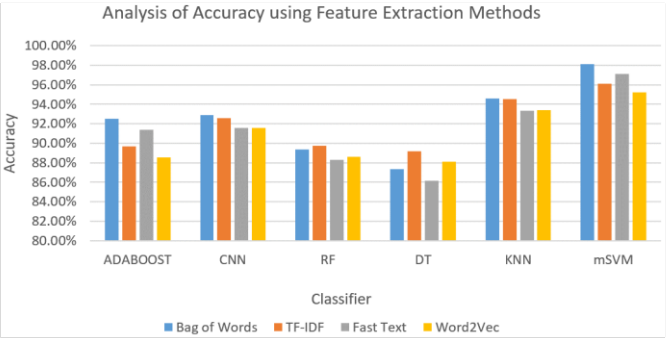

Figure 2. Analysis of accuracy using feature extraction methods.

The **Figure 3**, depicts the analysis of precision of proposed mSVM and traditional ML classifiers using various feature extraction methods. It is observed that the precision of mSVM using bag of words feature extraction method is 96.88% which is better as compared to other ML classifiers.

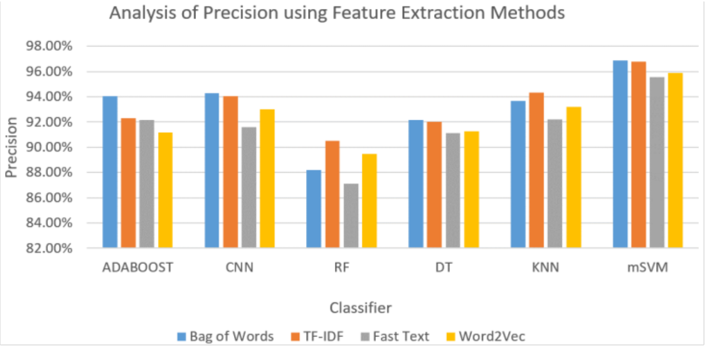

Figure 3. Analysis of precision using feature extraction methods.

The **Figure 4**, depicts the analysis of recall of proposed mSVM and traditional ML classifiers using various feature extraction methods. It is observed that the recall of mSVM using Bag of Words feature extraction method is 96.35% which is better as compared to other ML classifiers.

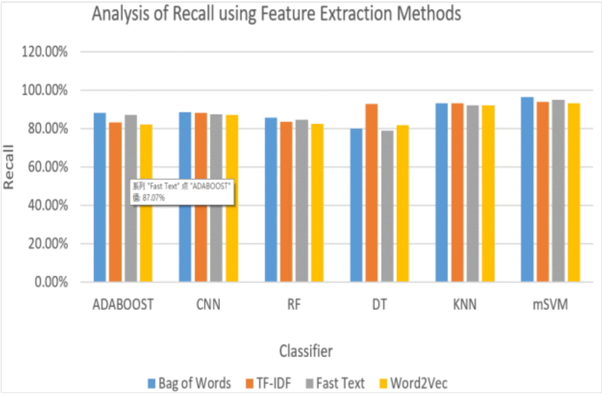

Figure 4. Analysis of recall using feature extraction methods.

The **Figure 5**, depicts the analysis of F1-score of proposed mSVM and traditional ML classifiers using various feature extraction methods. It is observed that the F1-score of mSVM using bag of words feature extraction method is 97.53% which is better as compared to other ML classifiers.

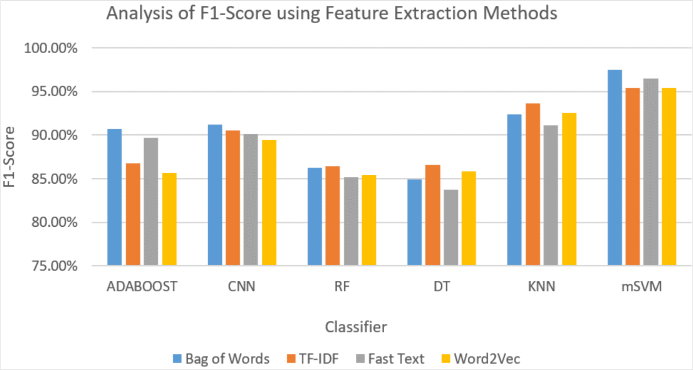

Figure 5. Analysis of F1-score using feature extraction methods.

Once the tweets are correctly classified as COVID-19 tweets, it is further categorized into three sentiments that is positive, negative and neural. The performance classification of this sentiments is performed using proposed mSVM and traditional ML classifiers which is depicted in the following **Table 2**.

In contrast to negative and neutral tweets, positive tweets have a higher chance of being correctly identified, as shown in the **Table 2**, which is the case for the majority of the classification methods. Proposed mSVM achieves 93% of accuracy rate for positive sentiment which better as compared to other ML classifiers.

Table 2. Performance analysis of each sentiments using proposed mSVM and traditional ML classifiers.

Classifiers	Sentiment	Precision	Recall	F1-score	Accuracy
Adaboost	Negative	0.78	0.8	0.79	0.81
	Neutral	0.62	0.83	0.71	0.67
	Positive	0.85	0.71	0.78	0.87
CNN	Negative	0.79	0.82	0.81	0.81
	Neutral	0.68	0.78	0.73	0.71
	Positive	0.83	0.76	0.78	0.86
RF	Negative	0.82	0.84	0.83	0.85
	Neutral	0.78	0.77	0.77	0.81
	Positive	0.83	0.81	0.82	0.86
DT	Negative	0.71	0.69	0.7	0.74
	Neutral	0.47	0.75	0.58	0.5
	Positive	0.76	0.59	0.66	0.79
KNN	Negative	0.82	0.82	0.82	0.85
	Neutral	0.76	0.76	0.75	0.78
	Positive	0.81	0.82	0.81	0.85
mSVM	Negative	0.89	0.92	0.9	0.92
	Neutral	0.84	0.83	0.84	0.86
	Positive	0.91	0.87	0.89	0.93

5. Conclusion and future scope

During the COVID-19 pandemic, tweets served as a source of data and could be relied upon as a trigger for virus surveillance models. The early reactions of public health agencies such as sending signals advance of outbreaks and offering advance warning before the pandemic develops, can be aided by analysing tweets, which can be helpful. In this work, the COVID-19 sentiments that were stated in tweets were analysed and categorised using mSVM, which is a dependable and useful source of data for researching people's behaviours and analysing vast amounts of data. Various feature extraction techniques namely bag of words, TF-IDF, Word2Vec and FastText was used to extract the relevant feature. Traditional ML classifiers and proposed mSVM model is trained using these feature extraction techniques. It is observed from the experimental findings that the performance accuracy of mSVM using bag of words method is 98.15% which is better than other ML classifiers. Once the tweets are correctly classified as COVID-19 tweets, it is further categorized into three sentiments that is positive, negative and neural. Proposed mSVM achieves 93% of accuracy rate for positive sentiment which better as compared to other ML classifiers. The first constraint of the strategy that was proposed is that it can only be applied to tweets written in the English language worldwide; however, in order to facilitate further research, the model can be modified to classify tweets written in languages other than English by including significant COVID-19-specific key phrases. For the purpose of comparison, analysis, and evaluation of this model, the data can also be incorporated from other platforms, the prevalence of which varies across different countries. One example of this type of platform is the Google trend dataset.

Author contributions

Conceptualization, CSL; methodology, CSL; software, CSL; formal analysis, CSL; investigation, CSL; resources, CSL; data curation, CSL; writing—original draft preparation, BSK; writing—review and editing, SS; visualization, BSK; project administration, SS. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

1. Zope T, Rajeswari K. Sentiment analysis of Covid-19 tweets using Twitter database—A global scenario. In: Proceedings of the 2022 IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA); 8–9 October 2022; Goa, India. pp. 27–30.
2. Chitra K, Tamilarasi A, Hemalatha S, et al. Sentiment analysis on Covid-19 vaccine. In: Proceedings of the 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC); 17–19 August 2022; Coimbatore, India. pp. 745–750.
3. Soomro ZT, Ilyas SHW, Yaqub U. Sentiment, count and cases: Analysis of Twitter discussions during COVID-19 pandemic. In: Proceedings of the 2020 7th International Conference on Behavioural and Social Computing (BESC); 5–7 November 2020; Bournemouth, United Kingdom. pp. 1–4.
4. Adamu H, Jiran MJBM, Gan KH, Samsudin NH. Text analytics on Twitter text-based public sentiment for Covid-19 vaccine: A machine learning approach. In: Proceedings of the 2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET); 13–15 September 2021; Kota Kinabalu, Malaysia. pp. 1–6.
5. Tareq A, Hewahi N. Sentiment analysis of tweets during COVID-19 pandemic using BLSTM. In: Proceedings of the 2021 International Conference on Data Analytics for Business and Industry (ICDABI; 25–26 October 2021; Sakheer, Bahrain. pp. 245–249.
6. Tao A, Qi K, Che D, et al. Comparison of media sources for COVID-19 by machine learning sentiment analysis. In: Proceedings of the 2021 International Symposium on Networks, Computers and Communications (ISNCC); 31 October 2021–2 November 2021; Dubai, United Arab Emirates. pp. 1–4.
7. Khan R, Rustam F, Kanwal K, et al. US based COVID-19 tweets sentiment analysis using TextBlob and supervised machine learning algorithms. In: Proceedings of the 2021 International Conference on Artificial Intelligence (ICAI); 5–7 April 2021; Islamabad, Pakistan. pp. 1–8.
8. Jannah HA, Hermawan D. Analysis of Indonesian society's perceptions of the COVID-19 vaccine in Youtube comments using machine learning algorithms. In: Proceedings of the 2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS); 7–8 September 2022; IPOH, Malaysia. pp. 72–77.
9. Patravali SD, Algur SP. Sentimental analysis of COVID-19 tweets using semantic approach. In: Proceedings of the 2022 3rd International Conference for Emerging Technology (INCET); 27–29 May 2022; Belgaum, India. pp. 1–4.
10. Sancoko SD, Diwandari S, Fachrie M. Ensemble learning for sentiment analysis on Twitter data related to Covid-19 preventions. In: Proceedings of the 2022 International Conference on Information Technology Research and Innovation (ICITRI); 10 November 2022; Jakarta, Indonesia. pp. 89–94.
11. Balaji TK, Bablani A, Sreeja SR. Opinion mining on COVID-19 vaccines in India using deep and machine learning approaches. In: Proceedings of the 2022 International Conference on Innovative Trends in Information Technology (ICITIIT); 12–13 February 2022; Kottayam, India. pp. 1–6.
12. Islam MN, Khan NI, Roy A, et al. Sentiment analysis of Bangladesh-specific COVID-19 tweets using deep neural network. In: Proceedings of the 2021 62nd International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS); 14–15 October 2021; Riga, Latvia. pp. 1–6.
13. Baker O, Liu J, Gosai M, Sitoula S. Twitter sentiment analysis using machine learning algorithms for COVID-19 outbreak in New Zealand. In: Proceedings of the 2021 IEEE 11th International Conference on System Engineering and Technology (ICSET); 6 November 2021; Shah Alam, Malaysia. pp. 286–291.
14. Nwafor E, Vaughan R, Kolimago C. Covid vaccine sentiment analysis by geographic region. In: Proceedings of the 2021 IEEE International Conference on Big Data (Big Data); 15–18 December 2021; Orlando, FL, USA. pp. 4401–4404.
15. Andhale S, Mane P, Vaingankar M, et al. Twitter sentiment analysis for COVID-19. In: Proceedings of the 2021 International Conference on Communication information and Computing Technology (ICCICT); 25–27 June 2021; Mumbai, India. pp. 1–12.
16. Mohsen A, Ali Y, Al-Sorori W, et al. A performance comparison of machine learning classifiers for Covid-19 Arabic Quarantine tweets sentiment analysis. In: Proceedings of the 2021 1st International Conference on

Emerging Smart Technologies and Applications (eSmarTA); 10–12 August 2021; Sana'a, Yemen. pp. 1–8.

17. Kumari KR, Gayathri T, Madhavi T. Machine learning technique with spider monkey optimization for COVID-19 sentiment analysis. In: Proceedings of the 2022 International Conference on Computing, Communication and Power Technology (IC3P); 7–8 January 2022; Visakhapatnam, India. pp. 303–307.
18. Senadhira KI, Rupasingha RAHM, Kumara BTGS. Sentiment analysis on Twitter data related to online learning during the Covid-19 pandemic. In: Proceedings of the 2022 International Research Conference on Smart Computing and Systems Engineering (SCSE); 1 September 2022; Colombo, Sri Lanka. pp. 131–136.
19. Aminuddin R, Bistamam MA, Ibrahim S, et al. A sentiment analysis framework on COVID-19 in major cities of Malaysia based on tweets using machine learning classification model. In: Proceedings of the 2021 IEEE 11th International Conference on System Engineering and Technology (ICSET); 6 November 2021; Shah Alam, Malaysia. pp. 25–30.
20. Dangi D, Dixit DK, Bhagat A, et al. Analyzing the sentiments by classifying the tweets based on COVID-19 using machine learning classifiers. In: Proceedings of the 2021 IEEE International Conference on Technology, Research, and Innovation for Betterment of Society (TRIBES); 17–19 December 2021; Raipur, India. pp. 1–6.
21. Pane SF, Prastya R, Putrada AG, et al. Reevaluating synthesizing sentiment analysis on COVID-19 fake news detection using spark dataframe. In: Proceedings of the 2022 International Conference on Information Technology Systems and Innovation (ICITSI); 8–9 November 2022; Bandung, Indonesia. pp. 269–274.
22. Guo R, Xu K. A large-scale analysis of COVID-19 Twitter dataset in a new phase of the pandemic. In: Proceedings of the 2022 IEEE 12th International Conference on Electronics Information and Emergency Communication (ICEIEC); 15–17 July 2022; Beijing, China, 2022. pp. 276–281.