Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

An Ensemble Machine Learning Techniques With Dolphin Swarm Algorithm For COVID-19 Sentiment Analysis

Vathsala.M.K.[1] , Sharmila Suttur C[2]

[1,2]Assistant Professor Dept Of ISE , MSRIT, Bangalore.

Abstract

The range of Covid-19 has raised concerns for the health of people all across the world. When faced with a catastrophe like a natural disaster, political turmoil, or terrorism, social scientists and psychologists want to know how individuals express their feelings and sentiments. There have been numerous psychological concerns caused by the COVID-19 epidemic, including depression in light of recent social changes and a lack of jobs. News and thoughts regarding it are rapidly being shared via social media. To make the best use of available resources, a realistic appraisal of the current situation is required. We use an ensemble ML technique to do sentiment analysis on Covid19 tweets in this study. To effectively deal with the current pandemic crisis, identification of Covid-19 attitudes from tweets is necessary. Text Blob is used to extract sentiments from the dataset after it has been cleaned up using pre processing procedures. Using our suggested feature set, the Dolphin Swarm Algorithm(DSA), we evaluated the performance of various machine learning classifiers. Positive, neutral, or negative tweets are all considered to be equal. Classifiers are judged based on their accuracy, precision, recall, and F1 score, among other metrics. A 90.51 percent accuracy score using our proposed concatenated feature set shows that Support Vector Machine (SVM) Classifiers outperform all other models. Compared to ML classifiers, the Multi-layer perceptron (MLP) has poor accuracy.

Keywords: COVID-19; Dolphin Swarm Algorithm; Machine Learning; Multi-Layer Perceptron; Sentiment Analysis.

Introduction

The Covid-19 outbreak has a negative influence on the economy [1]. On the 30th of January, 2020, the World Health Organization classified it as an epidemic [2]. Since then, it has expanded rapidly, resulting in numerous health problems and even fatalities. The death toll is expected to reach 636,633 by May 31st, 2020 [3]. There will be an ongoing virus outbreak for some time to come [4]. The amount of traffic on social networking sites has surged dramatically during the lock-down [5]. Covid-19 news was quickly disseminated via Twitter before it reached its rivals [6]. Due to personal opinions and bias, the vast majority of news is subjective, leading to intentionally false information, ambiguity, and negativity among people in social circles [7]. Meanwhile, academics are taking notice of this circumstance in order to do computer analysis in order to provide a comprehensive picture. This study employs supervised machine learning methods to analyze sentiment on the Covid-19 Twitter dataset.

Human emotional expression can be captured by analyzing content posted on social media sites like Twitter and Facebook. The media area is inundated by people's fears, numbers, facts, and general thoughts, and this data can disclose a lot about the general mood and temperament of the social population. Natural language processing (NLP) and artificial intelligence (AI) are becoming important in text analytics because of society's increased reliance on social media for info, rather than out-dated news sources, and the size of data that is now available. Social media is also used by many businesses to market their goods, brands, and services [9-10].

Natural language processing (NLP) and the use of it in social media analysis have grown at an astronomical rate. However, the difficulties in determining the intrinsic meaning of a document using NLP methods persist. Even the latest NLP tools have been demonstrated to be "susceptible to hostile texts" Take a look at the following tweets to see how people feel about COVID-19.

1) Best wishes for a prosperous new year. Wishing you all the best in the Year of the Rat, as well as plenty of cheese and immunity against the coronavirus.

2) The Corona Virus has infected 200 people, according to recent news reports. That's terrible.

3) A coronavirus' symptoms and geographic distribution are unknown.

4) SARS was a bad illness, but the new Coronavirus could be even worse, according to experts.

5) Mr. Peanut was killed by a coronavirus..

These tweets demonstrate how people feel about COVID-19 based on what they have said on Twitter. We need to focus on building efficient algorithmic techniques for automatically extracting writer's sentiment from text, which is where sentiment analysis comes into play. Research efforts are aimed at determining how people feel about a single statement, usually expressed in the form of a brief text post that is rife with subjectivity and uncertainty. One of the study's contributions is a discussion and comparison of text classification mechanisms based on ML approaches, which are commonly used in AI applications for natural language processing (NLP), but contextualized specifically in tweet sentiment classifications using ML in this research.

This research attempts to find out what people are saying about COVID-19 on Twitter, as well as the dynamics of their sentiment. Using Natural Language Processing (NLP), we discovered topics and detected sentiment in a novel labelled data set of COVID-19-related tweets. From March to September 2020, the Twitter API was used to collect streaming Twitter data, and the emotion found in the tweets was divided into three groups:

those with good, negative, and neutral messages. The data set is trained and validated using a variety of algorithmic models to deliver baselines for sentiment analysis linked to potential COVID-19 therapy discussions on Twitter. The best-performing model is designated for optimization and promotion after a final verification examination.

Everything after that is organized as follows. The related work is accessible in Section II. Section III discusses the research approach that was used to write this article. Section IV presents the findings, while Section V offers the conclusions.

2. Literature Review

Textual analytics and sentiment analysis, as well as machine learning (ML) and social media (Twitter, NLP) are all included in this part, which drew from a variety of academic areas. Data difficulties are increasing and must be solved with the use of ML approaches and strategic information features such as data rearrangement.

It's become common practice to utilize text analysis for a wide range of tasks, including email screening, irony identification, document organization, sentiment analysis, hate speech detection, and more. The use of Twitter data for emotional analysis has grown significantly. Due to the huge amount of data obtained and examined with a latent Dirichlet allocation (LDA) algorithm carried out by frequency-based filtering approaches, interesting findings were concealed in plain sight. Negative binomial and Poisson models [13] have both been used to examine tweet popularity. That research also examines the connections between the various subjects. They make use of a total of seven measures of difference. Using Kulliback–Leibler and Euclidean distances, the best user-based interactive approach-related subjects can be found. TAKE (time-aware knowledge extraction) was previously used in research

Comparable studies have utilized textual analysis to construct plans for human attribute identification, such in communication. Emotions from multilingual social media posts can be extracted using linguistic and psychological analysis [15]. Tracking Twitter data has also been used for epidemiological study, crisis scenario analysis, and other related tasks. Widener and Li [16] looked at the valances of healthy and unhealthy food emotion in different parts of the United States. The spatial distribution of studied tweets displayed that people in rural regions tweet less than those in cities and suburbs. It's also worth noting that the number of food tweets per capita was lower in smaller cities. Logistic regression found that tweets about unhealthy eating were more common in low-income locations.

Twitter data has also been put to use in healthcare sentiment analytics. A study by De Choudhury et al. [17] examines how new moms' moods and postnatal behavior alter after giving birth. It was found that data from Twitter can be effective in identifying mothers who are predisposed to postnatal despair by looking at language style, feeling, social network, and social engagement. New analytical frameworks have studied Twitter data linked to supply chain management (SCM), yielding important insights that improve SCM research and practice. [18]. Content analysis, sentiment analysis, and text mining were all performed on 22 399 SCM tweets, as well as descriptive analysis and network analytics. Carvaho et al. [19]

provide an effective platform known as MISNIS (intelligent Mining of Public Social Networks' influence on Society). To gather, store, manage, mine, and imagine Twitter data, this application was used.

According to Lopez et al. [20]'s study of social media research, government strategies towards the COVID-19 pandemic and popular debate on the pandemic issue were investigated. Text-mining of Twitter data in multiple languages from different nations is used to discover popular policy replies during the pandemic. They also employed text mining to show the epidemiology of COVID-19 using Bogota, Colombian press releases by Saire and Navarro [21]. There was a favorable correlation among the number of infected people and the quantity of tweets, they surmised According to Schild et al. [22], they examined 4Chan and Twitter to see how sinophobia spread throughout the pandemic. In a study by Kaila et al. [23], the top ten COVID-19 topics were derived from a random sample of 18000 tweets about the conference. The NRC sentiment lexicon was also used to compute the feelings, according to the study's authors. Han et al. [24] conducted a study on Chinese people's perceptions of COVID-19 emotion. COVID-19-related postings were divided into seven main categories, each of which had a total of 13 subcategories. According to Depoux et al. [25], social media panic spreads far more quickly than COVID-19 panic. As a result, specialists and appropriate authorities must notice and respond to such rumors, attitudes, and public conduct as quickly as feasible. Huang and Carley [26] have investigated the public sentiment and conversation on COVID-19 on Twitter and discovered that the posts of normal Twitter users are the maximum powerful. In contrast to the previous research, we used a new collection of data and examined what topics people were posting about the most. With the use of NLP, we were able to automatically detect sentiment.

2.1. Application to COVID-19 pandemic in India

India Due of the vast population and the large amount of densely populated cities, managing the COVID-19 pandemic was extremely difficult [27]. The first COVID-19 case was reported by India on January 30th, 2020, and the country went into lock-down mode, which was subsequently lifted. After the United States and Brazil (22nd March 2021), India had the third-highest number of confirmed cases with more than 11.6 million established illnesses and more than 160 thousand deaths. With almost 300,000 active cases, India was the world's 8th most affected country (prior to the second wave). Around the middle of September in 2020, the number of cases in India reached a significant peak of almost to 100,000 per day. This number dropped to around 11,000 per day by the end of January 2021, and has been steadily rising ever since. On the 22nd of March, India had 47 thousand new cases every day and was going towards a second high [28]. In March 2021, the instances began to rise rapidly.

In the first six months, Maharashtra (population 124 million), Delhi (population 19 million), and Tamil Nadu (population 8 million) dominated COVID-19 infections [29]. Mumbai is located in Maharashtra, a state with a population comparable to that of some of the world's most populous countries. Delhi cases decreased in the second part of the year, although they remained among the top eight states [30]. On a weekly basis in March 2021, Maharashtra continued to feature more than half of new cases and Delhi was able to contain the issue with less than a thousand instances per day with the highest infections. Since we are interested in both Maharashtra and Delhi, we will examine both states in depth.

We recognize that the structure we've proposed can be used elsewhere in the world, but we've chosen India as a case study to illustrate how well it works. COVID-19 data will be used in India in the last step, including sentiments from around the country and from two key states with COVID-19 cases. This reveals that both states had significant peaks surveyed by minor ones, but India had one huge peak around the middle of September 2020 with approximately to 97,000 new cases per day during the peak, according to the data.

3. Proposed Methodology

Our machine learning framework for sentiment analysis includes a number of steps, including tweet extraction, tweet pre processing, feature selection, model development, and ensemble machine learning training and prediction using selected COVID-19 data. This framework uses machine learning ensembles to perform sentiment analysis. Figure 1 depicts the suggested methodology's workflow.

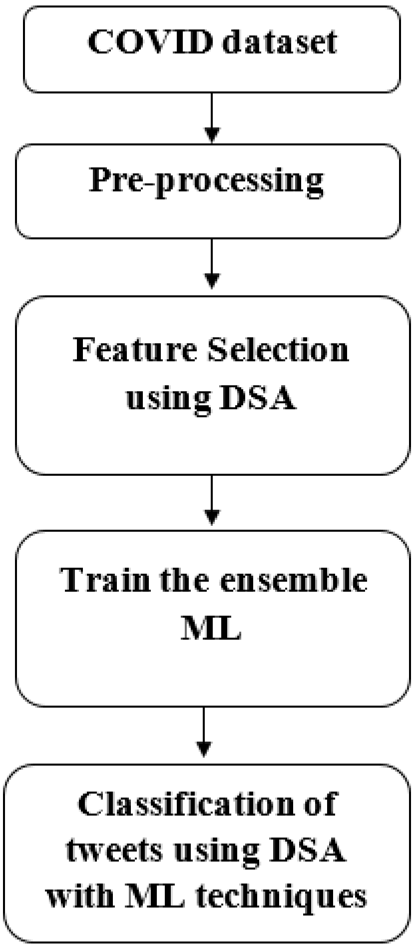

Figure 1: Working Flow of proposed methodology

3.1. Dataset Description:

As a starting point, we'll look at the COVID-19 dataset from a certain location. For this study, we looked at COVID-19 tweets from India and two states where the amount of COVID-19 cases was among the highest in 2020. (Maharashtra and Delhi). We used the COVID-19 India test dataset [31], which includes tweets about COVID-19 between March and September 2020. More than 150,000 Indian tweets make up this collection. We use this dataset to build two additional ones, one for Maharashtra (a state) and the other for Delhi (a union territory), each containing about 18,000 tweets. The Senwave COVID-19 dataset [32] has 11 different attitudes categorized by a panel of 50 specialists for 10,000 tweets worldwide during the COVID-19 pandemic in 2020, which we use hand-labeled sentiment data for. Despite the fact that "official report" is a topic, it is included in this dataset as a sentiment.

3.2. Data Pre-processing

As a result of this, each word must be translated into its associated Glo Ve embedding vector (300 dimensions in our case). For language models using sentiment analysis, Glo Ve embedding has demonstrated promising results [33]. There are several special qualities of Twitter language such as usernames, links, and hashtags that may not be significant in the classification process. Lemmatization is used to clean up the cleaned up data. Different word forms are transformed into fundamental words in this manner to create new ones. The input clause labels post before that. There were several unwelcome characters in the original data, such as punctuation and commas, as well as the full stops and semicolons that are commonly found in Twitter posts. Tokenization is the process of breaking up a long input sequence into smaller, more manageable chunks. Each word's aforementioned vector is then handed along to the various training models that have been developed.

3.3. Feature Selection

Techniques for identifying the most significant elements for classification are known as subset selection strategies or ranking approaches, respectively. When using the previous kind, you'll get back a subset of the most important features (qualities) for categorization. In order to classify the twitter data, characteristics are chosen depending on their specific classes. Adequate characteristics are only employed to classify the data throughout the classification process. Consequently, the features should be reduced in accordance with the requirement for sufficient information for categorization. In this study, the improved classification performance is provided by feature selection. The Dolphin Swarm Algorithm is used in the proposed technique to select the best feature (DSA).

3.3.1. DSA

Low classification accuracy is caused by useless data and an excessive amount of characteristics. Classifiers require appropriate data as input to tackle these problems. DSA-based feature selection method is used to carry out this approach.

A. Hunting Process

PSO has influenced the creation of a few meta-heuristic algorithms. DSA isn't an outlier in this regard. Dolphin swarms have been studied by Wu et al. and several behaviors such as echolocation, division of labor, collaboration, and the flow of information have been discovered in the dolphins. Dolphins use these behavioral characteristics to catch and eat their prey.

B. Basic Definitions

1) Dolphin

Each dolphin represents a plausible solution to the optimization problem in the DSA optimization process, which uses particles from the PSO. In dolphins are distinct as $DoI_i = [x_1, x_2, \ldots, x_D]^T (i = 1,2,..,N)$,Where$N$ mean odd dolphin and $x_j(j = 1,2, \ldots, N)$ mean the component.

2) Neighborhood Solution and Optimal Individual

The two significant meanings related to the DSA are the optimal separate solution and neighborhood solution. Moreover, for each $DoI_i = (i = 1,2,..,N)$,there are two significant factors which are $K_i(i = 1,2,..,N)$and $L_i(i = 1,2,..,N)$, where $L_i$as the optimal solution that DoI_i find in exclusive time, and K_i reperesents the optimal solution of the what DoI_i gets from others.

3) Fitness and Distance

Distances between Li and Ki (DLK i), Doli and Ki (DLK i), and Dol j have to be specified in order to use distance-based search algorithms for discrete-state systems (DSA). The following is the formula for expressing these three distances:

$$DD_{i,j} = \left|\left|Dol_i - Dol_j\right|\right|, i, j = 1,2, \ldots N, i \neq j. \qquad (1)$$

$$DK_i = ||Dol_i - K_i||, i = 1,2, \ldots N. \qquad (2)$$

$$DKL_i = ||L_i - K_i|| \qquad (3)$$

C. Critical Stages

1) Search Stage

When dolphins are on the prowl, they make a loud, piercing noise that can be heard for miles around. V i is defined as the sound component of each dimension in this study as M (i=1,2,...,M) in order to accurately depict the dolphin's hunting process for prey, while v j (j=1,2,...,D) signifies that each dimension has a amount of sounds and a direction attribute for the sound. The speed property of sound, 'speed,' also fulfills $|V_i = [v_1, v_2, \ldots, v_D](i = 1,2, \ldots, M)$.To keep dolphins out of the search phase, a maximum search time T1 has been imposed. There will always be a solution Xijt for the sound Vj that DoI i=(i=1,2,..,N) makes at time t in the interval 0 to T1. X_{ijt} is defined as follows:

$$X_{ijt} = Dol_i + Vj_t \qquad (4)$$

For **X**ijt, its fitness value Eijtis distinct as follows:

$E_{ijt} = Fitness(X_{ijt})$(5)

If

$E_{iab} = min_j = 12, \dots M;_{t=1,2,\dots,T_1} E_{ijt} = min_{j=1,2,\dots,M;t=1,2,\dots,T_1} Fitness(X_{ijt})$ (6)

Then Li of Doliis distinct as
$L_i = X_{iab}$(7)

If

$Fitness(L_i) < Fitness(K_i)$(8)

2) Reception Stage

After the call stage in DSA, there is a stage called reception. First, a close-up view of the reception area. With an N-order matrix called the "transmission time matrix," the information exchange between dolphins can be expressed, with the remainder of the time for the sound travelling from Doli to Dolj being represented by TSij. This reduction in all components $TSij\ (i = 1,2,\dots N; j = 1,2.\dots N)$ in the TS when dolphins enter the reception stage indicates that noises propagate on any component $TSij\ (i = 1,2,\dots N; j = 1,2.\dots N); j = 1,2.\dots N)$.

$TS_{ij} = 0$(9)

This demonstrates that Dolj is delivering sound to Doli, which Doli will then receive. And TSij will be replaced by a novel acquisition time called "maximum transmission time" in addition to that (T2). The associated sound will be received using this procedure. In addition, if you compare Ki to Kj,

$Fitness(K_i) > Fitness(K_i)$(10)

3) Call Stage

For K_i, K_j, and $TS_{i,j}$, if

$Fitness(K_i) > Fitness(K_i)$(11)

$TS_{i,j} > \left[\frac{DD_{i,j}}{A.Speed}\right]$(12)

Where A signifies as the acceleration. Then, $TS_{i,j}$is updated in the light of the subsequent equation:

$TS_{i,j} = \left[\frac{DD_{i,j}}{A.Speed}\right]$(13)

After all the $TS_{i,j}$ is uploaded, DSA enters the reception phase.

4) Predation Stage

During this time, each dolphin is actively searching for prey inside an area known as R2. The distance between the predation's sites and the dolphin's ideal neighborhood solution are determined by R2, as well. It is also necessary to calculate the maximum search radius, or R1, which may be found below:

$R_1 = T_1 \times speed$(14)

Then, Doli $(i = 1,2, \dots N)$ isas an illustration of how to calculate R2 and then update the dolphin's location.

(a) For Dol_i $(i = 1,2, \dots N)$, if

$DK_i \leq R_1$(15)

Next, R_2 is calculate on the basis of equation (16)

$R_2 = \left(1 - \frac{2}{e}\right) DK_i, e > 2$(16)

where e signifies as the radius reduction coefficient.

$NewDol_i = K_i + \frac{Dol_i - K_i}{DK_i} R2$(17)

(b) For $Dol_i$$(i = 1,2, \dots N)$, if

$DK_i > R_1$ (18)

And

$DK_i > DKL_i$ (19)

Next, R2 is designed on the basis of equation (20).

$$R_2 = \left(1 - \frac{\frac{DK_i}{Fitness(K_i)} + \frac{DK_i - DKL_i}{Fitness(L_i)}}{e.DK_i \frac{1}{Fitness(K_i)}}\right) DK_i e > 2 \quad (20)$$

After obtaining R2, Doli's new position new Dolican be obtained:

$$newDol_i = K_i + \frac{Random}{||Random||} R2 \quad (21)$$

(c) For Doli$(i = 1,2, \dots N)$, if it satiates equation (24) and

$DK_i < DKL_i$ (22)

Next, R2 is calculated on the basis of EQUATION (17).

$$R_2 = \left(1 - \frac{\frac{DK_i}{Fitness(K_i)} + \frac{DKL_i - DK_i}{Fitness(L_i)}}{e.DK_i \frac{1}{Fitness(K_i)}}\right) DK_i e > 2 \quad (23)$$

After obtaining R2, Doli's new position new Doliis got by EQUATION (21).

After Doli changes to the position new Doli, associating new Doli with Ki in terms of fitness, if

$Fitness(newDol_i) > Fitness(K_j)$ (24)

Then new Doli will take the place of Ki, or Ki will remain the same. Last but not least, if iteration's end condition is met, DSA moves on to the termination stage; if not, it moves on to the search stage. Once the best solution is obtained, the input is given to classifiers, which is described as follows:

3.4. Proposed Ensemble Classifiers

In this research study, three different machine learning techniques as ensemble technique is used for sentiment analysis, which is described as follows:

3.4.1. Multi-layer perceptron (MLP)

Most neural network architectures use MLP, notably the two-layered ones with input and output blocks coupled to hidden intermediary layers between them. The non-linear activation functions of the network's neurons differ for each model. Thus, a static linkage between the network input area and the output space can be performed. MLPs, on the other hand, frequently have time-delayed connections between hidden neurons and a layer of reference units. These blocks temporarily retain the output of hidden neurons (along with one weight) before re-inputting them. The buried neurons keep track of their previous activity in this manner, allowing the network to learn in small steps over time.

Assumption: The MLP provides a non-linear mapping between the input vector and its matching output vector This non-linear mapping has been maintained in a static setting for a substantial part of the research in this field. Static models, for example, can be used to model a wide range of real issues, such as character recognition. On the other hand, dynamic modeling is required for many practical activities, such as the forecasting of time series, vision, language, and motor control. The MLP architecture has been expanded several times to better comprehend situations of this type. Networks that allow information to flow from the output back into the input area, such as repeating networks or feedback networks, are a good example.

The regeneration technique, a tried-and-true method of simply adjusting the MLP coefficients, is one of the algorithm's major advantages. It's a term that's used to describe a network of neurons. During the training phase, the network's output is compared to a signal to see how well it functions. This is a supervised learning method. To update the weights in the output layer, a reprocessing method is employed. This type of algorithm has a steeper descent since it uses an error signal to illustrate how much the neural network's current output deviates from the desired output. The next step is to recalculate and fine-tune the hidden network inputs. A neural network analyses input signals at full strength, but the resulting error multiplies as the network trains and makes weight adjustments. Using a simple network, the back propagational gorithm formulation will be demonstrated. A linear output neuron's output signal is sent through a hidden layer. This network's output is compared to the target signal to see if there is a problem.

3.4.2. Neural Networks

Neuronal networks are collections of algorithms that imitate the way the human brain works to identify underlying links in data sets. A neural network is a group of neurons that can be either biological or synthetic in nature. Because neural networks are capable of adapting to changing input, they can produce the best possible results even when the output criteria aren't changed. With its roots in AI, neural networks are rapidly becoming popular in the creation of trading schemes. Figure 2 depicts the fundamental NN structure[34].

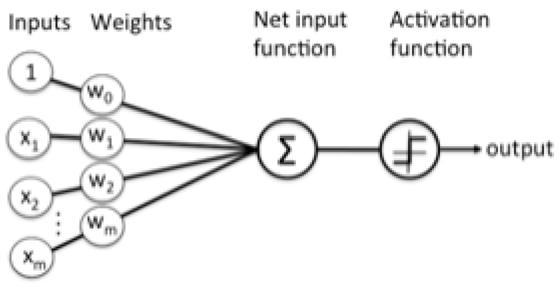

Figure 2: Basic Structure of Neural Network

There are no dependencies between the NN levels, thus each layer can have any number of nodes it wants. The bias node is the technical term for this arbitrary number of nodes. Bias nodes are always set to one by default. Nodes with bias resemble the offset in linear regression, which is written as $y = ax + b$, For ANN training success analysis, one can use bias values that change the activation function to the right or left, depending on the bias value. Nodes matching input features and output classes will be produced using a NN classifier. Input and output nodes will be matched. An input and an output node are often found in a NN when employed as a function approximation. However, there must be more hidden nodes than input nodes in the architecture. The same can be said for an artificial neural network. It's a three-layer system that works. The layer that receives input is referred to as the input layer. The input is processed by the layer that isn't visible. It's all over now, thanks to the output layer.

3.4.3. Support Vector Machine (SVM)

SVM is another popular modern machine learning approach. SVMs (supervised vector machines) in machine learning evaluate data for classification and regression analysis using supervised learning techniques. This technique, known as the kernel trick, allows SVMs to efficiently classify data in high-dimensional feature spaces while still achieving linear classification. In essence, it establishes boundaries between academic disciplines. In order to minimize the classification error, the margins are drawn with a distance between them and the classes that is as great as possible.

To increase the length of the border before reaching data points in a new category, you need to increase the margin of a linear classifier. You should go with the line that has the greatest difference between it and the other. Support Vectors are the data points that are on the outer edge of the distribution. Discovery of the hyper plane that best separates the two categories will be the next stage. A set of points is taken and parted utilizing a variety of application-specific mathematical algorithms using SVM to accomplish this. The positive and negative hyper planes can be deduced from this. (25-27) outlines the mathematical formula for locating the hyperplane:

$$(p.q) + r = +1 \ (positive\ labels) \quad (25)$$

$$(p.q) + r = \ -1 \ (negative\ labels) \quad (26)$$

$$(p.q) + r = \ 0 \ (Hyper\ labels) \quad (27)$$

From the equation above and using linear algebra we can discover the values of P and r. Thus, we get the model that contains the answers for $p\ and\ r$ with margin value of$2/2\sqrt{(k.k)}$, the margin is calculated as shown in Eq.(28):

$$Margin = 2/2\sqrt{(k.k)} \quad (28)$$

This model is used in support vector machines to classify new data. New data can be classified into several categories using the aforementioned solutions and the computed margin value.

4. Results and Discussion

Data on each disease is separated into training and testing datasets, with 80% of the data being used as training and the remaining 20% being used as testing in experimental work. open-source tool WEKA was used by the authors to gauge the performance of various machine learning methods..

4.1. Parameter Evaluation

As a result of measuring outcomes and results on a regular basis, the proposed system can generate accurate information about its efficacy. The process of reporting, collecting, and analyzing data on a group's or individual's performance is sometimes referred to as the performance measure. Table 1 shows the confusion metric used to evaluate classifiers for binary data, which may be explained as follows:

Table 1: Confusion Matrix

	Predicted Negative	**Predicted positive**
Actual Negative	True Negative (TN)	False Negative (FN)
Actual positive	False Positive (FP)	True Positive (TP)

The mathematical equation of accuracy, f-measure, precision, and recall are denoted in the Eq. (29), (30), (31), and (32).

$$Accuracy = \frac{TN+TP}{TP+TN+FN+FP} \times 1 \quad (29)$$

$$F-measure = \frac{2TP}{(2TP+FP+FN)} \times 1 \quad (30)$$

$$Precision = \frac{TP}{(FP+TP)} \times 1 \quad (31)$$

$$Recall = \frac{TP}{(FN+TP)} \times 1 \quad (32)$$

The experiments primarily aim to regulate the best method when compared with MLP, NN and SVM. Here the optimal features are selected by using DSA and here all techniques are tested with DSA and the result discussions are provided below:

4.2. Performance Analysis

In this section, the performance of all techniques are validated in terms of precision, recall and F-measure for three classes such as positive, negative and neutral in Table 2. The proposed techniques are provided in Table 3.

Table 2: Performance of proposed ensemble techniques for three classes

Methods	Parameters	Positive	Negative	Neutral
DSA-MLP	Precision	56	67	66
	Recall	59	71	76
	F-Measure	63	75	79
DSA-NN	Precision	62	73	69
	Recall	69	76	80
	F-Measure	71	78	81
DSA-SVM	Precision	82	76	86
	Recall	84	86	85
	F-Measure	86	84	83

The DSA-SVM achieved 82 percent precision for the positive class, 76 percent for the negative class, and 86 percent for the neutral class, whereas the greatest recall value for SVM with DSA is 85 percent for neutral, 86 percent for negative, and 84 percent for positive. SVM has an F-Measure of 86 percent for positive, 83 percent for neutral, and 84 percent for negative. When compared to SVM and NN approaches, MLP performs poorly on the negative, positive, and negative precision classes and achieves the best score of 79 percent on the neutral class. The NN technique achieved the greatest neutral value of 81% of F-measure, 80% of recall, and 69% of precision. SVM performs better because it is more effective in high-dimensional spaces, which is used with DSA for optimal features.

Table 3: Overall performance of ensemble techniques in terms of precision, recall and F-Measure

Methods	Precision	Recall	F-Measure
DSA-MLP	63	68.88	72.33
DSA-NN	68	75	76.66
DSA-SVM	81.33	85	84.33

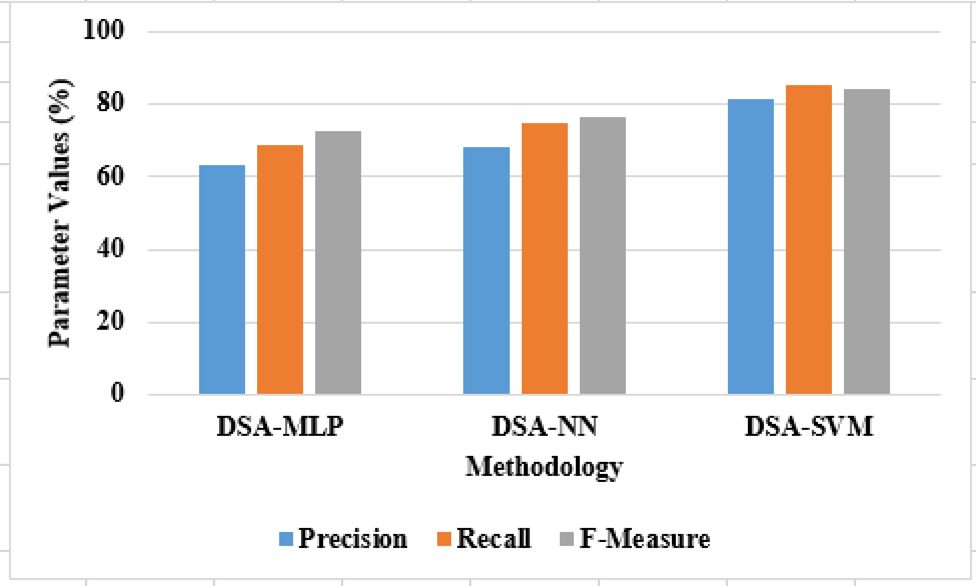

Figure 3: Performance of proposed ensemble technique with DSA for overall performance on various metrics

Table 3 clearly illustrates that the SVM achieved greater precision, recall, and F-Measure values when compared to other techniques such as NN and MLP. Using a weighting strategy and a clear margin of separation, the SVM achieved 81.33 percent precision, 85 percent recall, and 84.33 percent F-Measure. MLP provides less performance, with a precision value of 63 percent, a recall value of 68.88 percent, and an F-measure precision value of 72.33 percent. The NN attained a recall value of 75%, an F-measure of 76.66%, and a precision of 68%. The following section will discuss how ensemble approaches perform in terms of accuracy.

4.3. Performance Analysis of technique in terms of accuracy

In this section, the accuracy of proposed techniques with DSA and Table 4 represents the accuracy values of proposed methods.

Table 5: Accuracy Performance for proposed DT

Methods	Accuracy (%)
DSA-MLP	76.46
DSA-NN	83.97
DSA-SVM	90.51

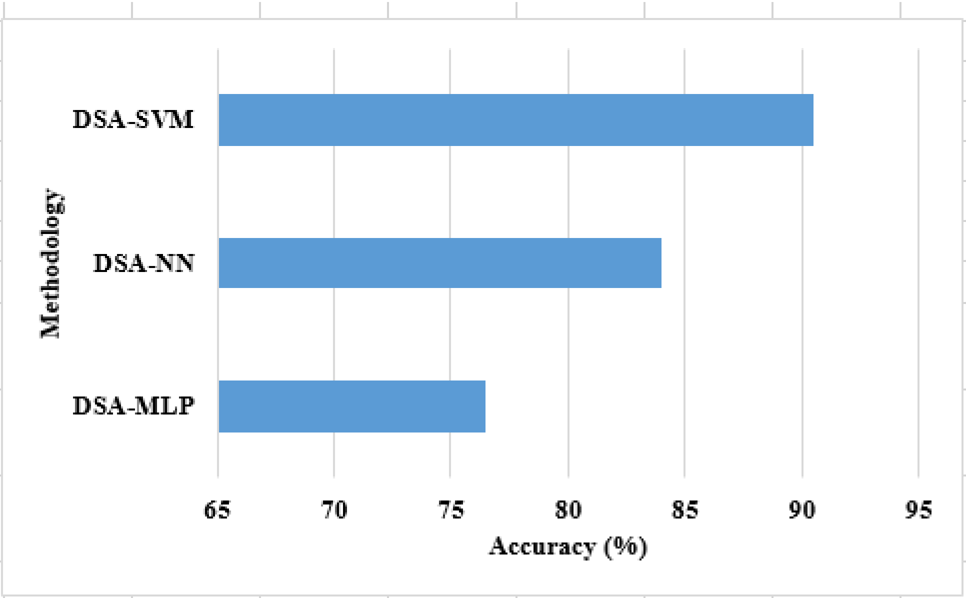

Figure 4: Performance of proposed ensemble technique in terms of Accuracy

According to the experiment results, the suggested SVM with DSA has a greater accuracy for classifying tweets from input datasets (i.e. 90.51 percent). When compared to other methods, the MLP had the lowest level of accuracy. The MLP method had an accuracy of 83.97 percent, however it did not pay attention to the amount of computing labor required. The input COVID dataset's testing results are compared to the training outcomes. This strategy outperforms previous methods like NN and MLP on the COVID dataset. Clearly, the amount of samples in the training and testing datasets influenced the accuracy of the predictions. Because the dataset is brand fresh, ML approaches performed worse when combined with DSA.

5. Conclusion

During the surge of COVID-19 infections, we conducted a study using unique machine learning algorithms for sentimental analysis. As a case study, we looked at tweets from India, specifically from Maharashtra and Delhi. There were 10,000 hand-labeled tweets in the COVID-19 dataset that we used to train our ML models. Towards the apex of new cases, the volume of tweets dropped significantly. Even while most people were upbeat, according to these forecasts, a sizable portion of the populace was disappointed in how the government handled the pandemic. With the increase in COVID-19 instances, the paper's key contribution is the structure for sentiment analysis in the population. Although our study ignores the effects of global COVID-19 news and statistics on other countries' overall opinion, this does have certain limitations. We can't use keyword selection for statistics seeking and Twitter filtering because it's not tied to a specific country. By looking at the tweets generated by certain communities or target societies, this research can be expanded in the future to determine the impact on public sentiment and attitude.

References

[1]. Iacus SM, Natale F, Santamaria C, Spyratos S, Vespe M. Estimating and projecting air passenger traffic during the COVID-19 coronavirus outbreak and its socio-economic impact. Safety Science. 2020; p.104791.

[2]. Guo YR, Cao QD, Hong ZS, Tan YY, Chen SD, Jin HJ, et al. The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak-an update on the status. Military Medical Research. 2020; 7(1):1–10. https://doi.org/10.1186/s40779-020-00240-0

[3]. worldometers. coronavirus death cases; 2020.

https://www.worldometers.info/coronavirus/coronavirusdeath-toll/.

[4]. Rustam F, Reshi AA, Mehmood A, Ullah S, On B, Aslam W, et al. COVID-19 Future Forecasting Using Supervised Machine Learning Models. IEEE Access.;.

[5]. Dubey AD. Twitter Sentiment Analysis during COVID19 Outbreak. Available at SSRN 3572023. 2020;.

[6]. Liu IL, Cheung CM, Lee MK. Understanding Twitter Usage: What Drive People Continue to Tweet. Pacis. 2010; 92:928–939.

[7]. Koohikamali M, Sidorova A. Information Re-Sharing on Social Network Sites in the Age of Fake News. Informing Science. 2017; 20.

[8] F. Barbieri and H. Saggion, “Automatic detection of irony and humour in Twitter,” in Proc. ICCC, 2014, pp. 155–162.

[9] W. He, H. Wu, G. Yan, V. Akula, and J. Shen, “A novel social media competitive analytics framework with sentiment benchmarks,” Inf. Manage., vol. 52, no. 7, pp. 801–812, Nov. 2015.

[10] M. J. Widener and W. Li, “Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US,” Appl. Geography, vol. 54, pp. 189–197, Oct. 2014.

[11] Z. Saeed et al., “What’s happening around the world? A survey and framework on event detection techniques on Twitter,” J. Grid Comput., vol. 17, no. 2, pp. 279–312, 2019.

[12] Z. Saeed, R. A. Abbasi, I. Razzak, O. Maqbool, A. Sadaf, and G. Xu, “Enhanced heartbeat graph for emerging event detection on Twitter using time series networks,” Expert Syst. Appl., vol. 136, pp. 115–132, Dec. 2019.

[13] J. Samuel, M. Garvey, and R. Kashyap, “That message went viral?! Exploratory analytics and sentiment analysis into the propagation of tweets,” 2020, arXiv:2004.09718.

[14] N. Ahmad and J. Siddique, “Personality assessment using Twitter tweets,” Procedia Comput. Sci., vol. 112, pp. 1964–1973, Sep. 2017.

[15] V. K. Jain, S. Kumar, and S. L. Fernandes, “Extraction of emotions from multilingual text using intelligent text processing and computational linguistics,” J. Comput. Sci., vol. 21, pp. 316–326, Jul. 2017.

[16] M. J. Widener and W. Li, “Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US,” Appl. Geography, vol. 54, pp. 189–197, Oct. 2014.

[17] M. De Choudhury, S. Counts, and E. Horvitz, “Predicting postpartum changes in emotion and behavior via social media,” in Proc. SIGCHI Conf. Hum. Factors Comput. Syst., Apr. 2013, pp. 3267–3276.

[18] B. K. Chae, “Insights from hashtag #supplychain and Twitter analytics: Considering Twitter and Twitter data for supply chain practice and research,” Int. J. Prod. Econ., vol. 165, pp. 247–259, Jul. 2015.

[19] J. P. Carvalho, H. Rosa, G. Brogueira, and F. Batista, “MISNIS: An intelligent platform for Twitter topic mining,” Expert Syst. Appl., vol. 89, pp. 374–388, Dec. 2017.

[20] C. E. Lopez, M. Vasu, and C. Gallemore, “Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset,” 2020, arXiv:2003.10359.

[21] J. E. C. Saire and R. C. Navarro, “What is the people posting about symptoms related to coronavirus in Bogota, Colombia?” 2020, arXiv:2003.11159.

[22] L. Schild, C. Ling, J. Blackburn, G. Stringhini, Y. Zhang, and S. Zannettou, “‘Go eat a bat, chang!’: An early look on the emergence of sinophobic behavior on Web communities in the face of COVID-19,” 2020, arXiv:2004.04046.

[23] D. P. Kaila et al., “Informational flow on Twitter–corona virus outbreak–topic modelling approach,” Int. J. Adv. Res. Eng. Technol., vol. 11, no. 3, pp. 128–134, 2020.

[24] X. Han, J. Wang, M. Zhang, and X. Wang, “Using social media to mine and analyze public opinion related to COVID-19 in China,” Int. J. Environ. Res. Public Health, vol. 17, no. 8, p. 2788, Apr. 2020.

[25] A. Depoux, S. Martin, E. Karafillakis, R. Preet, A. Wilder-Smith, and H. Larson, “The pandemic of social media panic travels faster than the COVID-19 outbreak,” J. Travel Med., vol. 27, no. 3, Apr. 2020, Art. no. taaa031.

[26] B. Huang and K. M. Carley, “Disinformation and misinformation on Twitter during the novel coronavirus outbreak,” 2020, arXiv:2006.04278.

[27]. Lancet T. India under COVID-19 lockdown. Lancet (London, England). 2020; 395(10233):1315.

[28]. Total Coronavirus Cases in India. Worldometer. Date last accessed: 26th, July, 2021: https://www.worldometers.info/coronavirus/country/india/

[29]. Unique identification authority of India; 2020. Date last accessed: 26th, July, 2021: https://uidai.gov.in/images/state-wise-aadhaar-saturation.pdf.

[30]. Chandra R, Jain A, Chauhan DS. Deep learning via LSTM models for COVID-19 infection forecasting in India. Co RR. 2021;abs/2101.11881.

[31] Lamsal R. Design and analysis of a large-scale COVID-19 tweets dataset. Applied Intelligence. 2020; p. 1–15.

[32] Yang Q, Alamro H, Albaradei S, Salhi A, Lv X, Ma C, et al. Sen Wave: Monitoring the Global Sentiments under the COVID-19 Pandemic. arXiv preprint arXiv:200610842. 2020.

[33] Pennington J, Socher R, Manning CD. Glo Ve: Global Vectors for Word Representation. In: Empirical Methods in Natural Language Processing (EMNLP); 2014. p. 1532–1543.

[34] Mahesh, B., 2020. Machine Learning Algorithms-A Review. International Journal of Science and Research (IJSR).[Internet], 9, pp.381-386.