1 INTRODUCTION:

Data Science is focused on interdisciplinary domains and useful for taking decisions. Effective Vaccines are needed to save lots of lives throughout the worldwide epidemics such as COVID19. The community looks to COVID-19 vaccination progression must be considered sensibly in directive to know the users sentiments and fears to it. To know more about exact information about covid-19 vaccines are from who are taken the vacci nation and they are express their opinions. In this research article studied and understands the advantages of social media. Now a Social media has become an important tool for gaining insights about any domain. At the time of Covid-19 pandemic social media applications are playing a key role in users thoughts on various topics sharing. About Vaccination side effects and results confusion is one of the serious issues in realizing herd immunity and suppressing the COVID-19 epidemic. To consider this approach our focus on analyze user opinions on COVID-19 vaccination process. The world face a main corona virus epidemic from the year 2019. The virus infects fast through various ways. All nations lock peoples to avoid the virus. Vaccinations, including Covaxin, Covishield, Pfizer, Moderna, SputnikV have been permitted. This research article, tweet analysis is based on people's opinions about official covid-19 vaccines on social media Twitter. Datasets collected, Covaxin, Covishield, Pfizer, Moderna, SputnikV. These tweets are preprocessed using Machine learning techniques.
In this research article studied the users opinion on Pfizer, Modern, AstraZeneca and Johnson & Johnson. The total posts in each nation for time period of month of Jan 2020 to Apr 2020, May 2020 to Aug 2020 and Sept 2020 to Dec 2020 was plot. The use of opinion Analysis impacts on each domain like product analysis, Recommendation system, prediction on healthcare and analytics.
After declaration of vaccination and governments announces the policy about vac cination. More peoples hesitates about its impact and side effects. In the month of Nov 9, 2020, when the vaccination drive starts and many people are reacting on social media about their effectiveness .

2 Survey of Literature:

The sentiment analysis is techniques it is used to identifying the users expressions and for that Sentiment Analysis, Machine learning, Natural language processing are popular techniques are effectively used. In this research article here perform the twitter appli cation management and collect real time hashtags discussion on covid-19 vaccination. The twitter general public data collection and preprocessing techniques are applied. Our investigation detected that unigram Sentiment Analysis for all five datasets. Lexicons are used Bing Liu and Sentiment140 are used for interpreting the data. The study is completed on the tweets which are related to the COVID-19 vaccination. Also focused on closest the users approaches of the COVID-19 vaccination process on twitter as a social media platform using machine learning. Maximum of the described sentiments that debated the vaccines effectiveness, security, and the distribution plans of Govern ments and the plans to safe the dosages for their people. This research analyzed the users opinions since the vaccination drives was started. Logistic Regression classifiers shows highest correctness was 97.3%, SVM model that shows correctness of 96.26% and MNB model shows correctness of 88%..
In the raw dataset 16 attributes where collected and then apply preprocessing for removal of noise and outliers. The users required towards distinguish whether present vaccine can stop the spread of the COVID19.
The results indicate that the Machine learning classification techniques for product

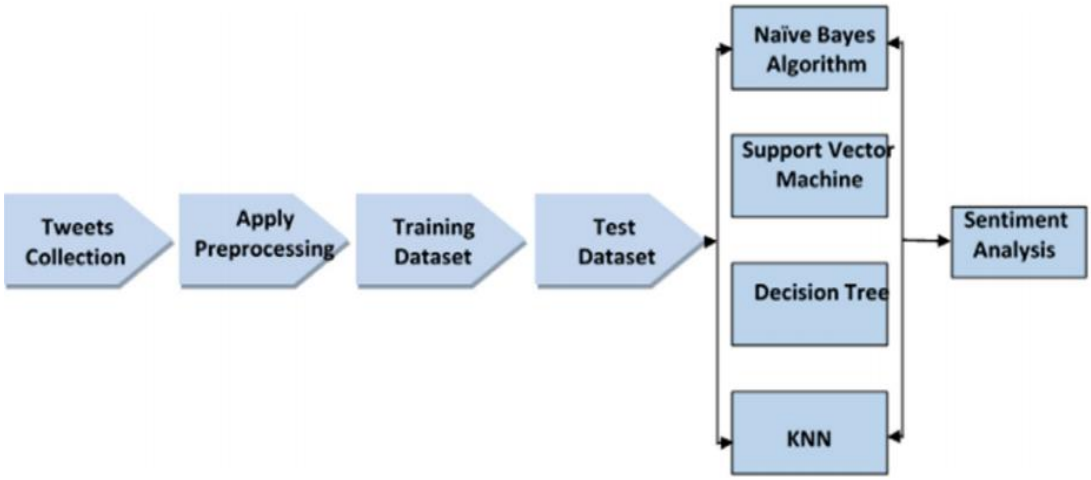
reviews has achieved the maximum classification correctness in comparison with Classification techniques.
Vaccine uncertainty slowed due to the few reasons protection, doubt about political forces driving the COVID-19 epidemic, a deficiency of information about the vaccine, confusing content of social media

3 Contribution of Work:

Through data Analysis provides the popularity of users on Vaccination and they may able to understand the sentiments about vaccines.
To Show the Effective use of Machine learning and Sentiment Analysis in Medical domain. Our study emphasized the vital essential for communicating public services with the humanity from diverse social and instructive families in direction to rise the vaccination consciousness and authenticate analysis. Study the various datasets related to Covid- 19 Vaccination and identifying the insights in the form of Positive opinions, Negative opinions and Neutral opinions.

4 Methodology：

The machine learning common approach is designed it performs the Tweet collection, Tweet Preprocessing, Train the Dataset, Test the Dataset and apply the classifiers and obtain the results



Sr. No	Dataset	Duration	Total Size
1	Covaxin	01 June to August 2021	5000
2	Covishield	01 June to August 2021	5000
3	Pfizer	01 June to August 2021	5000
4	Moderna	01 June to August 2021	3500
5	SputnikV	01 June to August 2021	5000

4.1 Data Preprocessing ：

Data preprocessing needs because in the proposed work we get raw data with 16 attributes of twitter datasets. Data preprocessing removed the noisy and duplicated data and convert into the quality data. Data Preprocessing Removing the URLs, Data Filtering, Removing Special Characters, Removal of Retweets, Usernames, Remove Punctuations and sym bols, Usage of Web links, Hashtags, Tokenization, Exclamation and question marks, Letter Repetition, Negations.

4.2 Train Datasets:

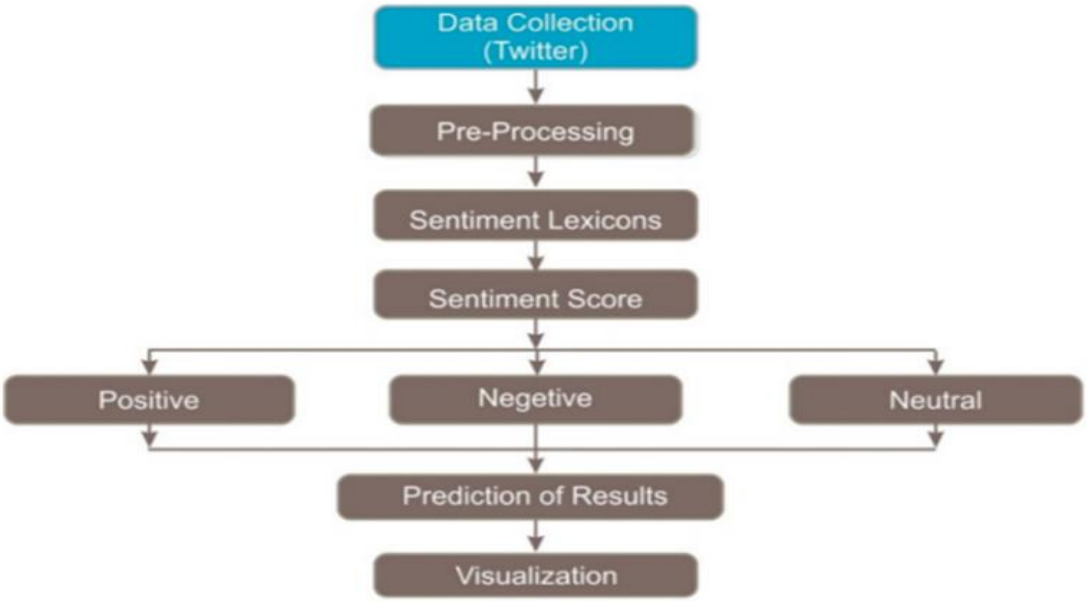
In this research work combining the machine learning and lexicon based technique. The Features are created. The Machine Learning classifiers applied for data classifications improved the accuracy with 70:30 Ratio where 70% Train Data and 30% Test Data.

4.3 Test Datasets:

In this research here test following twitter dataset. Covaxin, Covishield, Pfizer, Moderna, SputnikV.

4.4 Lexicon Based Approach:

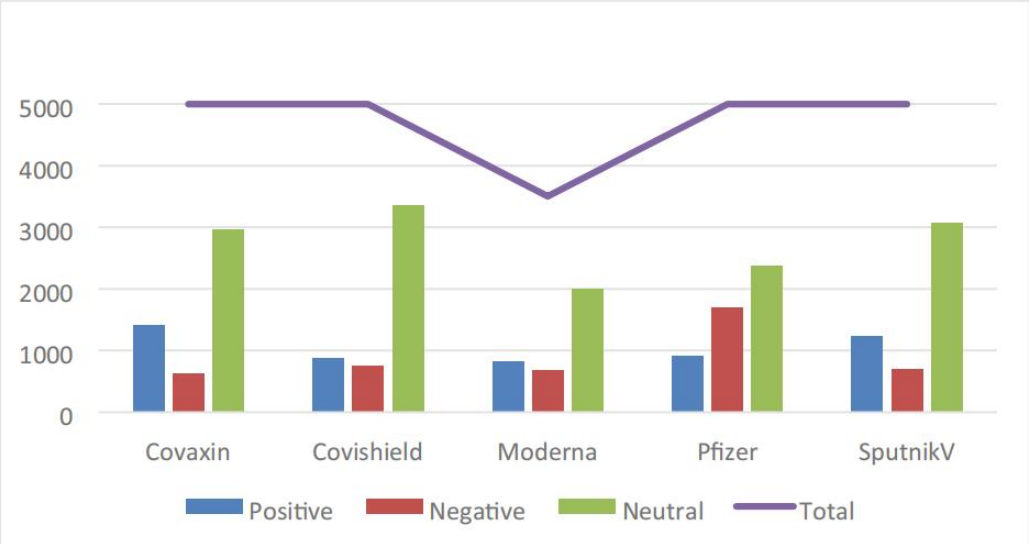
In this research work used Stanford University sentiment lexicons, it contains total 1,600,000 from that half of tweets are from Positive and remaining half from Negative Lexicons. Tweets collected and prepared by Stanford University, where the tweets are categorized based on an occurrence of positive and negative score. Also used Bing Liu sentiment lexicons. In the Sentiment Lexicon method, the Bing Liu Dataset having 2,006 and 4,783 Positive and Negative lexicons respectively in this dataset around 6789 words.



Name of Vaccination	Positive	Negative	Neutral	Total
Covaxin	1407	625	2968	5000
Covishield	881	752	3367	5000
Moderna	829	675	1996	3500
Pfizer	914	1703	2383	5000
SputnikV	1231	705	3064	5000

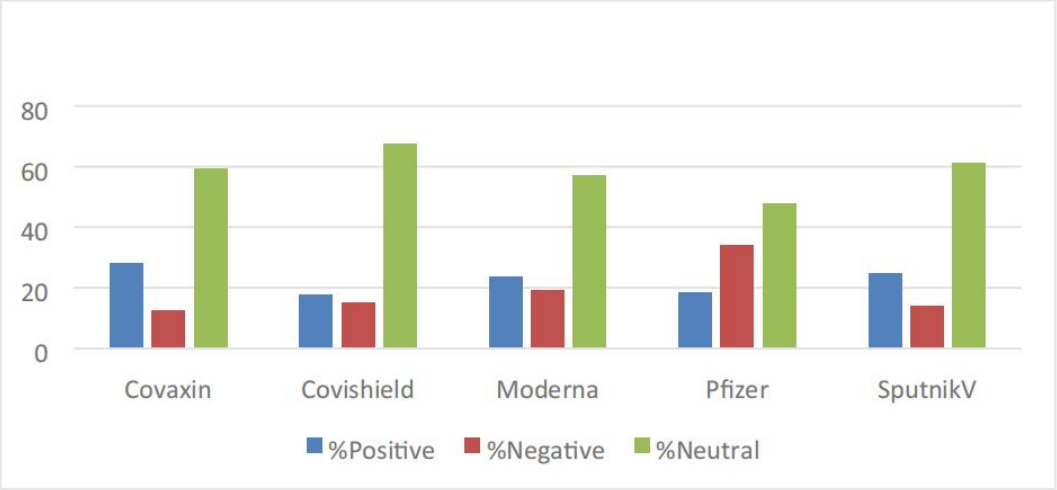
4.6 Sentiment Score:

The Key insights Opinions of users to produce opinion of each Sentence. Sentiment Score = sum (pos.matches) – sum(neg.matches)
In the above Opinion Mining of Covaxin Twitter Dataset total of 5000 tweets are extracted, in which 1407 tweets are categorized as positive, 625 as negative and remaining 2968 are categorized as neutral tweets. In Covishield Dataset total of 5000 tweets are extracted, in which 881 tweets are categorized as positive, 752 as negative and remaining 3367 are categorized as neutral tweets. In Moderna Dataset total of 3500 tweets are extracted, in which 829 tweets are categorized as positive, 675 as negative and remaining 1996 are categorized as neutral tweets. In Pfizer Dataset total of 5000 tweets are extracted, in which 914 tweets are classified as positive, 1703 as negative and remaining 2383 are categorized as neutral tweets. In SputnikV Dataset total of 5000 tweets are extracted, in which 1231 tweets are classified as positive, 705 as negative and remaining 3064 are classified as neutral tweets. The overall opinion analysis of vaccination is most of the data shows the neutral opinion about vaccination but in another approach of peoples when

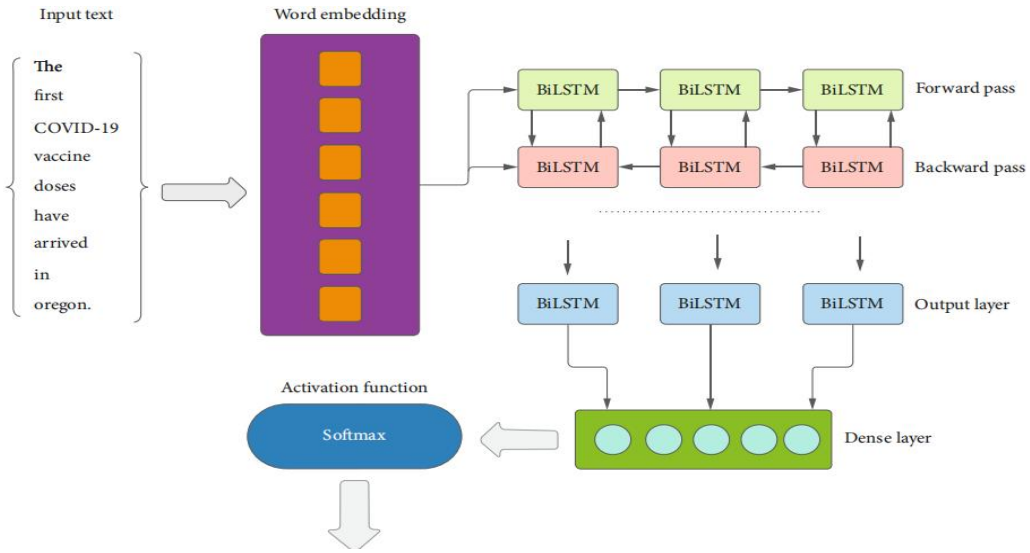


Name of Vaccination	%Positive	%Negative	%Neutral
Covaxin	28.14	12.5	59.36
Covishield	17.62	15.04	67.34
Moderna	23.68	19.28	57.02
Pfizer	18.28	34.06	47.66
SputnikV	24.62	14.1	61.28

we consider here most positive expressions of peoples on Covaxin and most negative expressions on Pfizer according to twitter data (Fig. 3 and Table 3).
In the above Opinion Mining of Covaxin Twitter Dataset total of 5000 tweets are extracted, in which 28.14% tweets are categorized as positive, 12.5% as negative and remaining 59.36% are categorized as neutral tweets. In Covishield Dataset total of 5000 tweets are extracted, in which 17.62% tweets are categorized as positive, 15.04% as negative and remaining 67.34% are classified as neutral tweets. In Moderna Dataset total of 3500 tweets are extracted, in which 23.68% tweets are categorized as positive, 19.28% as negative and remaining 57.02% are categorized as neutral tweets. In Pfizer Dataset total of 5000 tweets are extracted, in which 18.28% tweets are categorized as positive, 34.06% as negative and remaining 47.66% are categorized as neutral tweets. In SputnikV Dataset total of 5000 tweets are extracted, in which 24.62% tweets are categorized as positive, 14.1% as negative and remaining 61.28% are categorized as neutral tweets.



A series of equations describe the gates of LSTM [19]. Before describing the equation, it is necessary to first com prehend some of the variables used in these calculations. The sigmoid activation function σis used, the weight matrix is Wi , the previous LSTM block's output is represented by



problematic solving process result of tree Classifier applied by following expression

$$Info(D) = \quad pi \log2(pi)$$

K-**Nearest Neighbor**

L-

K-Nearest Neighbor (K-NN) algorithm is a technique for classifying data based on nearest class that are neighboring to the data. For calculating space is called as Euclidean Distance.

$$D(x, p) = \sqrt{(x - p)}$$

ht−1, and the preference for the corresponding gates is repre sented bybi . Finally, the existing time stamp input is xt, and the input gate is it: $i_t = \sigma$

(Wi *ht−1, xt ½+ bi)

The data that can be given to the cell were chosen using this equation. The forget gate f t in Equation (2) decides which data from the input side of the previous memory should be ignored:

f t = σ (Wi ht−1, xt ½ + bi)ð Þ

In Equation (3), tanh normalizes the values in the range between -1 and 1, where C is the candidate for the cell state at the timestamp, which controls the cellðtÞ updates:

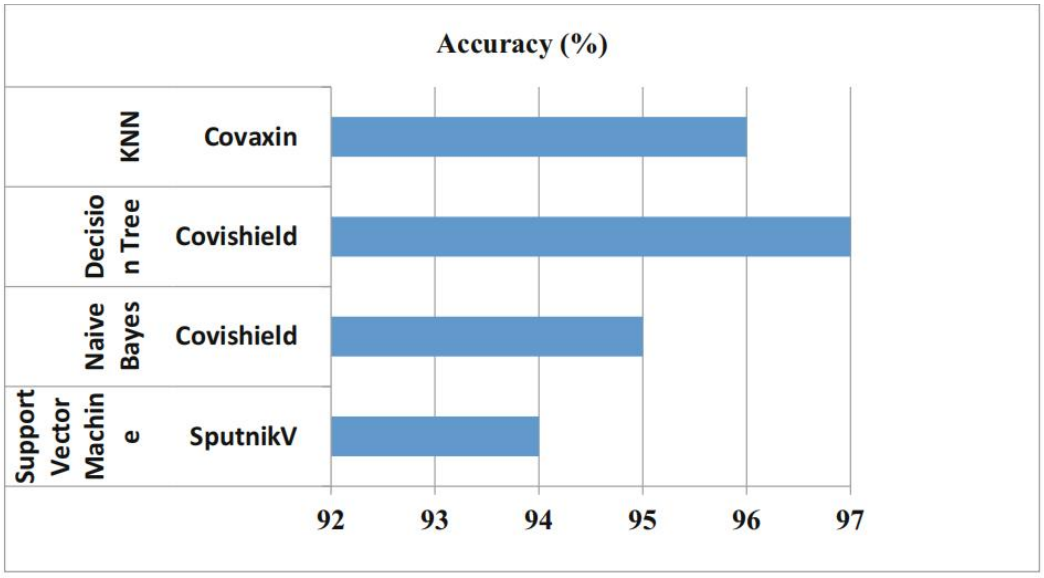
C = tanh(Wc ∗ ht−1, xt ½ + bc),

Ct = f t × Ct−1 + it × C:

The output layer (ot) upgrades both the hidden layer ht−1 and the output layer according to Equation (4):

ot = σ(Wo ∗ ht−1, xt ½+ bo) ,

ht = ot × tanh Ct



Other Performance Metrics:

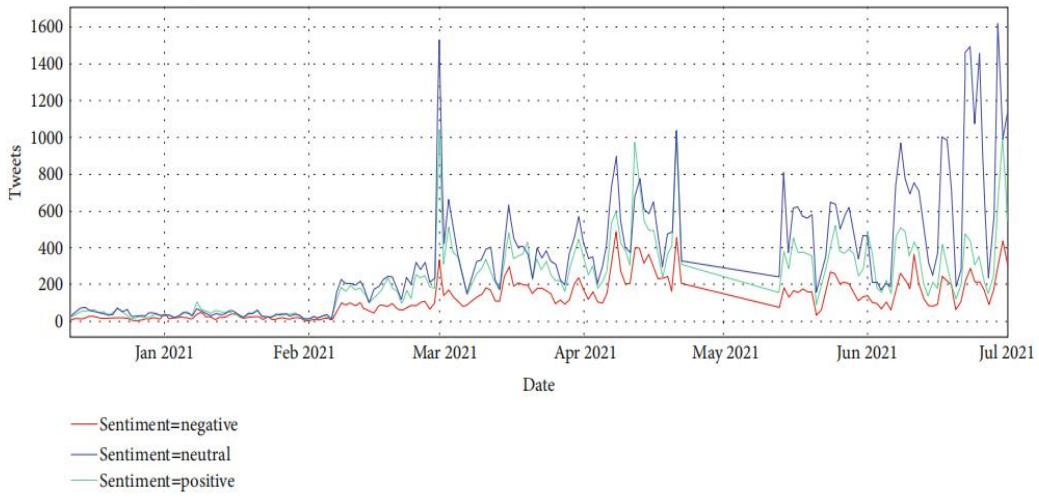
To evaluate the perfor mance of the model based on different metrics, this study used precision, recall, F1-score, and a confusion matrix with different values such as true positive (TP), false positive (FP), true negative (TN), and false negative (FN).

Precision: this describes the performance of the model on the test data. It shows the number of models predicted correctly from all the positive classes:

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall: the percentage of total relevant results accurately classified by the algorithm is referred to as the recall:

$$\text{Recall} = \frac{TP}{TP+FN}$$



Conclusions:

In this paper, we collected data from twitter and then apply preprocessing for data exploration, classification. Data Preprocessing involves the Removing URLs, DataFiltering, Removing Special Characters, Removal of Retweets, Usernames, Remove Punctuations and symbols, Usage of Web links, Hashtags, Tokenization, Exclamation and question marks, Letter Repetition, Negations.

Machine learning classifier used and studied the comparative analysis between KNN, Support Vector Machines, Naïve Bayes, Decision Tree algorithms for data classification.

Finding shows that Decision Tree classifier for Covishield dataset has achieved the highest 97% accuracy with compared to Naïve Bayes, Support Vector Machine, KNN classification methods. Support Vector Machine has lowest Accuracy with 94% for SputnikV.

COVID-19 Vaccination dataset wise machine learning model evaluation perfor mance studied and got highest and lowest results of Machine learning classifiers. The Support Vector Machine SputnikV dataset got highest accuracy with 94% and Cov ishield dataset got lowest accuracy with 89%, The Naïve Bayes got highest accuracy for Covishield dataset with 95% and lowest accuracy with 87% for Moderna dataset, The Decision tree got highest accuracy for Covishield dataset with 97% and lowest accuracy with 88% for Pfizer dataset, The KNN got highest accuracy for Covaxin dataset with 96% and lowest accuracy with 88% for SputnikV dataset.

In Lexicon Based approached Sentiment polarity classification here total 23500 tweets taken for result analysis and predict the vaccination opinions on SputnikV, Covishield, Covishield, Covaxin, Pfizer datasets. Overall here identify the Neutral opinions on Vaccinations. In other side when we focused on positive and negative opinions here Covaxin is more positive compare with all other vaccination datasets according twitter discussion of users insights and negative opinions on Pfizer vaccination datasets.