# SENTIMENT ANALYSIS ON TWITTER DATA OF COVID-19 VACCINES USING MACHINE LEARNING

**A Thesis Submitted**

**In Partial Fulfillment of the Requirements**

**For the degree of**

## MASTER OF TECHNOLOGY

in

**Software Engineering**

**by**

**AMRITA MISHRA**

**1190449001**

**Under the Supervision of**

**Mr. Mohd. Saif Wajid**

**Associate Professor**

**Mrs. Upasana Dugal**

**Associate Professor**

**Babu Banarasi Das University, Lucknow**

**to the**

**School of Engineering**

**BABU BANARASI DAS UNIVERSITY**
**Lucknow,**
**June, 2021**

# CERTIFICATE

It is certified that the work contained in this thesis entitled **“Sentiment Analysis on Twitter data of COVID-19 Vaccines Using Machine Learning”,** by **AMRITA MISHRA** (Roll no 1190449001), for the award of **Master of Technology** from Babu Banarasi Das University has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

Signature upasana

Mr. Mohd. Saif Wajid
Mrs. Upasana Dugal
Department of CSE
BBD University, Lucknow
Date: 23. 05.2021

# ABSTRACT

With the development of the Web, the Twitter has emerged as the most useful platform where people can share, upload, disseminate their feelings about goods or services, governmental issues, economic movements, recent developments, and a plethora of such social interactions. Twitter data is now accepted worldwide as important data source and a center for accessing information about diseases, opinion mining and digital marketing.

The year 2020 has witnessed worldwide spread of COVID-19 disease. A lot of research studies based upon the causes and sentiments of people about disease have been carried out. However, on 12 August 2021, Russia developed the first registered vaccine Sputnik V accompanied by U.S developed vaccine Moderna on 17 December, 2021 and India's based Covaxin on 03 January, 2021. The problem of analyzing the sentiments about the vaccines and calculating the accuracy of sentiments is feebly carried out.

In this study, lexicon based sentiment mining methods have been applied to know the opinion of people about different vaccines which are Sputnik V, Moderna and Covaxin and the results are validated using Machine Learning.

For this study, Twitter developer account is created which fetched 2000 tweets of individuals in Python anaconda environment. The collected data was subjected through Lexicon based sentiment polarity computation techniques. The obtained sentiment score for tweets of Vaccines is individually validated by Naïve Bayes Algorithm.

This study proposes Lexicon based approach for sentiment analysis and Naïve Based algorithm for predicting the accuracy of results from lexicon based technique, a new framework for data analysis. This technique overcomes the problem of limited architecture for comparison of sentiments about vaccines. For the implementation of the proposal, Python Anaconda environment has been used. Some Word Processing techniques like Tokenization, Stemming, Word count, N-grams are also carried out.

The performance of the proposed architecture, evaluated using Naïve Based Algorithm gave an accuracy of 79 % for Sputnik V, 70% for Moderna and 78% for Covaxin. The result has shown that Covaxin yields maximum positive sentiments than Sputnik V and Moderna.

# ACKNOWLEDGEMENTS

I hereby declare that the thesis entitled **“Sentiment Analysis on Twitter data of COVID-19 Vaccines Using Machine Learning”** submitted by me in the Partial Fulfillment of the requirements for the award of the degree of Master of Technology ( CS-SE) of BBDU (BABU BANARASI DAS UNIVERSITY) is record of my own work carried under the supervision and guidance of **Mr. Mohd. Saif Wajid and Mrs. Upasana Dugal .**

I am also very grateful and thankful to our Head Sir Dr. Praveen Kumar Shukla who guided us in right way.

To the best of my Knowledge this thesis project has not been submitted to Babu Banarasi Das university or any other university or Institute for the award of my degree.

**AMRITA MISHRA**

# TABLE OF CONTENT

# LIST OF FIGURES

# LIST OF TABLE

# LIST OF ABBREVIATION

| | |
|---|---|
| API | Application Programming Interface |
| NLP | Natural Language Processing |
| NLTK | Natural Language Tool Kit |
| ML | Machine Learning |
| CSV | Comma Separated Values |
| P | Positive |
| N | Negative |
| Ne | Neutral |
| SRS | Software Requirement Specification |

# Chapter 1

## INTRODUCTION

### 1.1Introduction

Today the world is a Machine Dependency era. Well formed systems for information exchange from peer to peer or B2B are established. The need of the hour is to ensure that besides navigating the datasoil, customer's sentiments are evaluated. The correct assessment of user sentiments proves to be highlighting feature in winning or losing the product's name and growth in market. Earlier the information and feedback exchange systems were file and paper based which was accessible by limited people. However, today social media like Twitter serves as major platforms where users freely expresses their opinions and it is accessible within remote areas. The developers can even analyze the tweets based upon selective geographic locations and form conclusions on regional basis.

Sentiment analysis is the area which deals with judgments, responses as well as feelings, which is generated from texts, being extensively used in fields like data mining, web mining, and social media analytics because sentiments are the most essential characteristics to judge the human behavior. Customer sentiments can be found in tweets, comments and reviews. For example Reviews delivered by customer on online sites after purchasing the product, property or visiting the hotels. Sentiment Analysis plays a vital role in Data Science as it brings forward a computational study about the diverse opinions .The calculative study provides a platform to derive the true meaning of customer opinion whether positive or negative or neutral.

What more that sometimes even if customer gives bias comments like , "The A.C. Compressor is working well but costly one. " can be examined by machine to draw accurate conclusion. Furthermore dwelling into study of Sentiment Analysis involves descriptive study of SA via Machine Learning and Lexicon based approaches. Some basic types of SA are Fine grained, Emotion based, Aspect based and Multilingual based. Fine grained is used when polarity precision is important for a business. Example, Very positive = 5 stars and Very Negative = 1 star. Emotion based aims at detecting emotions, like happiness, frustration, anger, sadness, and so on while Aspect Based analyze sentiments of texts, let’s say product reviews to know which particular aspects has positive, neutral, or negative way. Multilingual Analysis involves a lot of preprocessing and resources e.g. translated corpora or noise detection algorithms are the There has been a lot of past research on different strategies to use the web innovation to expand the advantages of clients and in addition organizations in the commercial center.

COVID-19 is an infectious disease caused by recently found virus known as SARS-CoV-2(Severe Acute Respiratory Syndrome). Its outbreak is beyond the previous observations of this virus and is thus considered pandemic by World Health Organization .Some of the major vaccines developed for preventive and emergency use in COVID-19 are Sputnik V, Pfizer BioNTech, Moderna and Covaxin.

Sputnik V is developed by Russia on 12 August, 2020 by Gamaleya Research Institute (GMI) in collaboration with Russian Defense Ministry is the first registered vaccine. U.S also developed Pfizer BioNTech and Moderna on 11 December, 2020 and 17 December, 2020 respectively. On December 11, the Food and Drug Administration (FDA) issued an Emergency Use Authorization (EUA) for emergency use of Pfizer-BioNTech for prevention of coronavirus disease 2019 (COVID-19) for individuals. Moderna vaccine is used for active immunization to prevent COVID-19 caused by severe acute respiratory syndrome coronavirus 2(SARS-CoV-2) in individuals 18 years of age or above. Covaxin, India’s indigenous COVID-19 vaccine by Bharat Biotech is developed in collaboration with the Indian Council of Medical Research (ICMR) and National Institute of Virology (NIV).Covaxin has been granted approval for emergency restricted use in India on January 3, 2021.

## 1.2Proposed Problem Statement

Now days many techniques is available to study about COVID-19 sentiments but very few techniques provide the comparative analysis and accuracy prediction for its vaccines.

## 1.3Objective

Objective of this thesis are:

- Analysis of public information from social media could yield interesting results and insights into the world of public opinions.
- Sentiment analysis through Natural Language Processing (NLP).
- Classifying opinions in text into categories like "positive" or "negative" or "neutral" using lexicon method.
- It is need to be thinking about accuracy of data using Machine learning.
- Now days, Twitter data increases in lots of way, so it is important that we concern about gathering correct opinion with suitable techniques and better accuracy.
- Data visualization is also essential to gain complete information about public opinion.

## 1.4Motivation

Nowadays one of the most interesting topics is analysis of people sentiments which is an attractive study in order to draw the future planning and creation of new ideas. Generally, existing works focus on sentiments about the electronic gadgets and politics. This thesis refers to the gathering of opinions about the vaccines for infectious disease which is transmitted through contact and by small droplets produced when people cough, sneeze or talk, it is now becoming a source of depression, stress and anxiety because of misleading information posted on social media. Micro-blogging is the easy way of accessing writing of sentiments through the twitter posts (tweets) because this is the era of technology and smart phones. The most interesting in sentiments analysis we can predict everything like political views, interesting brand of electronics, sports, boutiques, hotel and resorts, stock exchange, movies, beautiful countries nature wise and technology wise, biggest events and many more.

## 1.5 Scope of work

Most previous methods have performed sentiment analysis in python using its inbuilt libraries. These libraries analysis the sentiments using the Dictionary based approach and threshold concept. The approach we claim here differs from these existing approaches in that we propose a a framework which can classify the sentiment polarity in the tweets and also predict the accuracy of polarity results. In order to further improve the generalization performance, we draw the Pie chart Polarity distribution of all the vaccines method to calculate the polarity percentage with higher efficiency.

## 1.6 Thesis Organization

In this thesis chapter 1 contains the introduction, chapter 2 contains the literature review details, chapter 3 contains the details about feature extraction, chapter 4 contains the classification details, chapter 5 shows architecture details, chapter 6 describe the result and chapter 7  provide conclusion of this thesis.

# CHAPTER 2

# LITERATURE SURVEY

Sentiment Analysis has been of avid interest to researchers lately. A lot of work has been put into it and there is a vast domain of its applications. Gaurav Bhatt et al.,2016,[21] have performed Sentiment Analysis over Educational institutions Using Twitter Dataset of IIT,NIT and AIIMS Colleges in India with SVM, Naïve Bayes and ANN algorithms and accuracy of 89.6%.The area of Neural Networks has been investigated for performing sentiment analysis on benchmark dataset consisting of online product reviews.

Bespalov, Bi,Qi and Shokoufandeh , 2015, [20] carried out binary classification on Amazon and TripAdvisor dataset using Perceptron classifier and obtained one of the lowest error rates among their experiments of 7.59 and 7.37 on the two datasets respectively. Researchers have also been working upon prediction of accuracy of tested datasets using Machine Learning Algorithms.

Kanakraj and Guddeti , 2015, [3] used Natural Language Processing techniques for Sentiment Analysis and compared Machine Learning Methods and Ensemble Methods to improve on the accuracy of classification. Shahheidari, Dong and Bin Daud, 2013, [4] used a Naïve Bayes Classifier for classification and tested it for news, finance, job, movies and sportstaking into consideration Data Mining on basis of two emoticons (☺ and ☹). Prediction Of Election Results is another domain in which massive population expresses opinion over Social Networks. Rincy Jose and Varghese S Chooralil, 2015 [7] have used Twitter Data with Classifier Ensemble Approaches with accuracy of 71.48% in predicting election results. Rincy Jose, et. al, 2015 [9] have also predicted election results with Word Sense Disambiguition with accuracy of 78.6% .

Mohd. Saif Wajid and et al. 2017, [8] have used Sentiment Analysis Based on A.I. Over Big Data. They have introduced the methodology for creating user recommended data group (Big data) by elaborating a matrix for user recommended data group for big data which is then reduced by dimension reduction technique. Neethu M.S and Rajasree R, 2013 [5] used twitter post on electronic products, compared the accuracy between different Machine Learning

Algorithmn and further improved accuracy by replacing repeated character with two occurrences, including a slang dictionary and taking emoticons into consideration.

Jotheeswaran and Koteeswaran, 2015 [6] performed binary classification on the IMDB dataset by employing a Multi-Layer Perceptron Neural Network and using Decision Tree -Based Feature Ranking for feature extraction and a hybrid algorithmn (based on Differential Evolution and Genetic Algorithm)for weight training, thereby obtaining a maximum classification accuracy of 83.25%.

Laszlo and Attila, 2020 [31] have used fresh scraped data collections over the Recurrent Neural Networks  to determine what emotional manifestations occurred in given time interval in COVID-19 .The Sentiment Analysis helps in monitoring area based upon the opinion raised in different territories. As Novel Corona Pneumonia (NCP) reports a major warning to the international population health, cooperation is required by all the countries to combat it [54]. The media and other social media channels should ethically present the relevant and 235 correct reports to increase motivation among the general public instead of presenting biased information; such coverage may only serve to divide individuals and stoke fear.

Sural et al. [37] demonstrated that Trait Emotional Intelligence (TEI) is directly associated with Problematic Social Media Use (PSMU) and indirectly associated with motives of presenting a popular side and passing time. This result interpreted as those 240 individuals who are lower on TEI use PSMU as a coping strategy to deal with their real life troubles. A sample of 444 individuals aged between 18 to 43 years and having an active social media account were taken. The statistical analysis was carried out with SPSS 23 and AMOS 23 software and for path analysis, maximum likelihood estimation was used. Direct and Indirect relationships were analyzed by using the 245 bootstrapping method with 5000 bootstrap samples and 95% bias-corrected confidence intervals.

Hornung et al. [38] presented the relationship between EI and Facebook® use. A sample of 105 individuals (60 female and 44 male participants) was taken. The sample's average age is 25.53 years and average number of Facebook® Journal Pre-proof Journal Pre-prooffriends is 365.97.

First, the measurement model of the directly observed indicators 250 was assessed from the questionnaire (such as Use and four EI dimensions) and second, the latent variable scores for the four EI dimensions were assessed by the higher-order constructs as well as the structural model. The research stated that the relationship between EI and Facebook® use is very positive for the younger group and very negative for the older group. A younger group uses Facebook® more and is 255 accustomed to it as they grow up with social media. They possibly develop their EI through or along with social media use and through social media networks. Herodotou et al. [39] examined the role of trait Emotional Intelligence (trait EI) for play and frequency of gaming over a sample of 1051 young adult US/European gamers (96% males and 4% females).

Chen et al. [40] investigated the influencing factors of the 260 three types of user social media intentions. A sample of 502 social media users including 52% females and 48% males aged between 21 to 35 years was taken for data analysis through online survey. Partial Least Squares (PLS) analysis and estimation were performed in two phases. The first phase conducted reliability and validity analysis, whereas the second phase estimated and verified the path 265 coefficients of the structural model which concluded that social media marketing activities have a significant influence on three types of intentions i.e. continuance intention, participation intention and purchase intention.

Kim et al. [41] examined the relationships between narcissism, the Big 5 personality traits, the need for popularity, the need to belong, and various types of selfie posting behaviors—posting solo selfies, 270 selfies with a group, and editing selfies. A sample of 260 participants was taken. The study found that selfie behaviors are associated with various personality traits and psychological needs. The results of this study suggested that a range of interpersonal motivations as well as egocentric traits underlie individuals' selfie activities. Depoux et al. [42] discussed the necessary measures to be taken in order to combat the 275 pandemic of social media panic. The paper emphasized to use social media wisely so that proper use of digital technologies can overcome the problem and help people balance their mental health using supportive measures from health ministry and solidarity during the stressful period of quarantine.

Merchant et al. [43] explained the importance of disseminating information through various social media platforms spend assets on preparing, we discover it is more essential to pick a decent encoder—which can frequently be a basic feed forward non-linearity. Our outcomes remember best in class execution for both CIFAR and NORB.

Santosh K. Divvala et.al., 2012, [44] The Deformable Parts Model (DPM) has as of late developed as an extremely valuable and well-known apparatus for handling the intra-classification variety issue in object identification. In this paper, we sum up the vital experiences from our exact investigation of the significant components comprising this identifier. All the more explicitly, we study the connection between the function of deformable parts and the combination model segments inside this indicator, and comprehend their relative significance. To start with, we find that by expanding the quantity of parts, and exchanging the instatement venture from their perspective proportion, left-right flipping heuristics to appearance based bunching, extensive improvement in execution is acquired. In any case, more intriguingly, we saw that with these new segments, the part misshapenings would now be able to be killed, yet getting outcomes that are nearly comparable to the first DPM indicator.

NavneetDalal, et. al., 2005,[45] We study the subject of capabilities for hearty visual item acknowledgment, receiving straight SVM based human identification as an experiment. In the wake of looking into existing edge and inclination based descriptors, we show tentatively that lattices of Histograms of Oriented Gradient (HOG) descriptors fundamentally beat existing capabilities for human identification. We study the impact of each phase of the calculation on execution, presuming that one-scale inclinations, one direction binning, generally coarse spatial binning, and top notch neighborhood contrast standardization in covering descriptor blocks are exceptionally significant for good outcomes. The new methodology gives close ideal division on the first MIT person on foot information base, so we present an additionally testing dataset containing more than 1800 commented on human pictures with a huge scope of posture varieties and foundations.

**Table 1**: – Relevant works pertaining to the pandemic

| S.No. | Motivation and aim of the work | Datasets used | Methods used in the work and Obtained Accuracy | Limitations |
|---|---|---|---|---|
| 32 | To trail the growth of panic amongst Twitter® users based on a specific keyword . | Tweets from Twitter® API was extracted using R language. Around 9 lakhs tweets were analysed in this work. | Naïve Bayes and Logistic Regression classifiers have been used. The accuracy for shorter tweets was obtained as 91% and 74%, respectively. | The paper analysed the sentiments based on a single keyword tracking based on only the fear of the people based in USA. Further aspects barring geographical constraints could also be explored. |
| 33 | To analyse the sentiments of Indians post lockdown imposed by the government | Datasets was collected using Twitter® application interface by R. Approximately 24 thousand tweets were used by extracting from the handles #IndiaLockdown and #IndiafightsCorona within the time span of 25th to 28th March, | Analysis was done by the help of software R and by using WordCloud only | Very few tweets were considered of a particular country. The study portrayed that Indians took the strategy of the government positively on imposing the lockdown. |

|  |  |  |  |  |
|---|---|---|---|---|
|  |  | 2020 |  |  |
| 34 | To trail the out the economic, political and health related impact on the people as envisaged through CoronaTracker website. | Data was collected from the CoronaTracker website between 22nd January and 3rd March, 2020 | Susceptible Exposed Infectious Recovered (SIER) model was used to predict the outbreak of the | The news extracted for analysing dates up to the beginning of March, 2020. Data was |
| 35 | To extract an exact idea by detecting the primary topics tweeted by netizens related to COVID-19 pandemic | Around 1,68,000 tweets were considered for grouping them into various topics | Tweets were analysed by unigrams and bigrams and were influenced by Dirichlet allocation for helping in topic modeling. | Though twelve topics were identified to be posted by the users during the span of February ’20 to March ’20, but they emphasized on the sentiments of the topics. These were mostly related to health care issues |
| 36 | To study how the Chinese Weibo users | Weibo messages were used from available 17,865 | Paired sample ttest by SPSS (Statistical Product and Service Solutions) | As Weibo users are mostly youth, the results were anticipated |

| | | | | |
|---|---|---|---|---|
| | were affected emotionally on and after 20th January, 2020 | active users of this platform within the time span of 13th to 26th January, 2020. | was used to measure the emotional features of the Weibo users | to be biased. The study shows how a rise in negative sentiments occur post the outbreak of the pandemic in the minds of the youth. |

# CHAPTER 3

# MATERIALS AND METHODS

This workc depicts about the prerequisites and focuses on major modules to design and implement the sentiment analysis on Twitter data. It determines the equipments and programming prerequisite thatcare needed for software keeping in mind the end cgoal, to run thecapplication appropriately. Thec Software Requirementc Specification (SRS) is clarified incpoint of interest, which incorporates outline of this exposition and cadditionally the functional andcnon-practical necessity of this thesis.

## 3.1 General Description

Most previous methods have designed sentiment analysis architecture using different inbuilt python library. The analyzer automatically calculates the polarity. The approach we claim here differs from these existing approaches in that we propose a framework where the sentiment scores obtained by analyzers can be validated by Machine Learning Algorithm. In order to further improve the data visualization for tweets of vaccines, we performed the Pie chart and Word Cloud generation for all .

### 3.1.1 Users Perspective

The characteristic of this task work is to give information about handling the Twitter data in terms of gaining access to Twitter developer account credentials, connecting the twitter with API in Python , sharing information in csv files and provide brief insight about the vaccines.

## 3.2 Feasibility Study

Believability is the determination of paying little respect to whether an undertaking justifies action. The framework followed in building their strength is called acceptability Study, these kind of study if a task could and ought to be taken.

Three key thoughts included in the likelihood examination are:

Technical Feasibility

Economic Feasibility

Operational Feasibility

### 3.2.1 Technical Feasibility

Here it is considered with determining hardware and programming, this will effective fulfill the client necessity the specialized requires of the framework should shift significantly yet may incorporate

The office to create yields in a specified time.

Reaction time underparticular states.

Capacity to deal with a particular segment of exchange at a specific pace.

### 3.2.2 Economic Feasibility

Budgetary examination is the often used system for assessing the feasibility of a projected structure. This is more usually acknowledged as cost/favorable position examination. The method is to center the focal points and trusts are typical casing a projected structure and a difference them and charges. These points of interest surpass costs; a choice is engaged to diagram and realize the system will must be prepared if there is to have a probability of being embraced. There is a consistent attempt that upgrades in exactness at all time of the system life cycle.

### 3.2.3 Operational Feasibility

It is for the most part identified with human association and supporting angles. The focuses are considered:

What alterations will be carried through the framework?

What authoritative shapes are dispersed?

What new aptitudes will be needed?

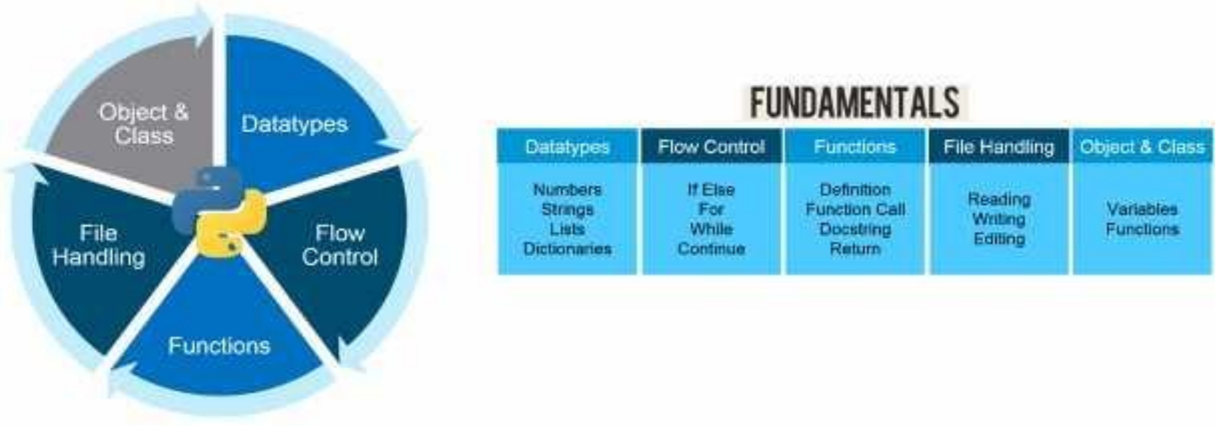
Do the current framework employee’s individuals have these aptitudes?

If not, would they be able to be prepared over the span of time?

## 3.3Technology used

### 3.3.1 PYTHON

Python is one of the most fast growing programing language in terms of number of developers. Developers mostly use python because it is one of the platform for easiest and fast coding and compilation. Python has huge number of libraries (scientific computing and data sciences) and many big companies use python, such asGoogle, Yahoo, YouTube, Dropbox and NASA. Python also supports machine learning, GUI, software developing and web developing, which are some of the reasons, it is used in this thesis. Python is general purpose, interpreter, object oriented and high level language. Python is also multi-paradigms programing language like functional, imperative, object oriented and reflective language. Python consists of different syntax and semantics such as Indentation, Statement and control flow, Expression, Methods, Typing and mathematics.



FUNDAMENTALS

| Datatypes | Flow Control | Functions | File Handling | Object & Class |
|---|---|---|---|---|
| Numbers<br>Strings<br>Lists<br>Dictionaries | If Else<br>For<br>While<br>Continue | Definition<br>Function Call<br>Docstring<br>Return | Reading<br>Writing<br>Editing | Variables<br>Functions |

Figure 1: Python Fundamentals

Python is a general-purpose high level programming language that is widely used in data science and for producing deep learning algorithms. Python and its libraries like Numpy, Scipy, Pandas, Matplotlib; frameworks like Theano, TensorFlow, Keras  for deep learning.

There are three different typing ducks, dynamic and gradual. Duck typing isan object can be used for a particular purpose. With normal typing, suitability is determinedby an object type (python 3.7.1, 2018). Dynamic is the set of rule properties that are called types to the various constructs of computer program such variable, function, expression or module. Programming language can effectively detect program errors at compile-time.

Moreover, some recent studies have indicated that the use of types can lead to significant enhancement of program performance at run-time(Xi, el, 1998). Siek and Jeremy said that the gradual typing is a type system in which some variables and expression may be given type and the correctness. Gradual typing allows software developers to choose either type paradigm as appropriate, from within a single language.

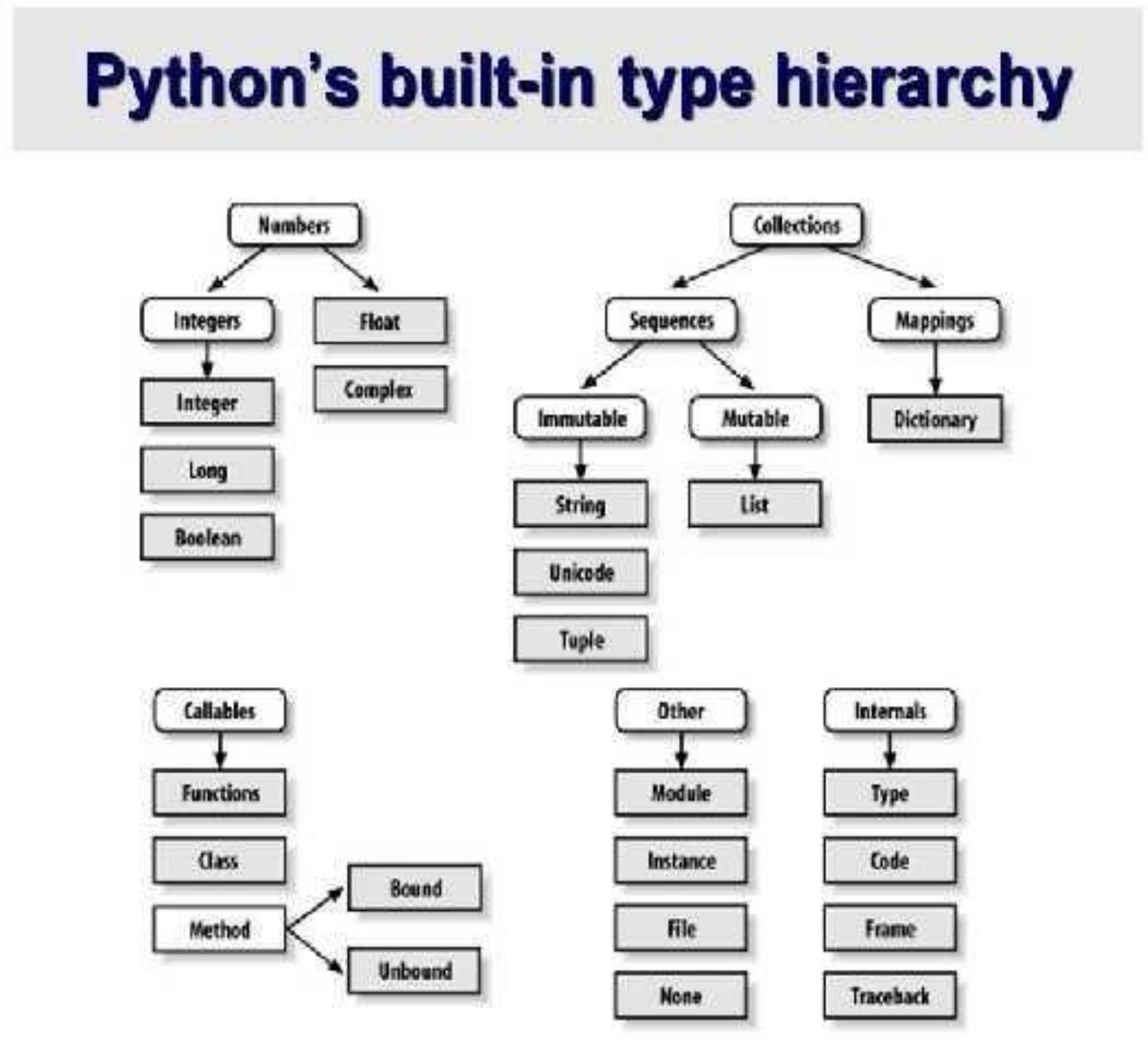


Figure 2:Python Hierarchy

### 3.3.2 ANACONDA

Anaconda is a distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. The distribution includes data-science packages suitable for Windows, Linux, and mac OS. It is developed and maintained by Anaconda, Inc., which was founded by Peter Wang and Travis Oliphant in 2012. As an Anaconda, Inc. product, it is also known as Anaconda Distribution or Anaconda Individual Edition, while other products from the company are Anaconda Team Edition and Anaconda Enterprise Edition, both of which are not free.

Package versions in Anaconda are managed by the package management system conda. This package manager was spun out as a separate open-source package as it ended up being useful on its own and for other things than Python. There is also a small, bootstrap version of Anaconda called Miniconda, which includes only conda, Python, the packages they depend on, and a small number of other packages (Wikipedia).

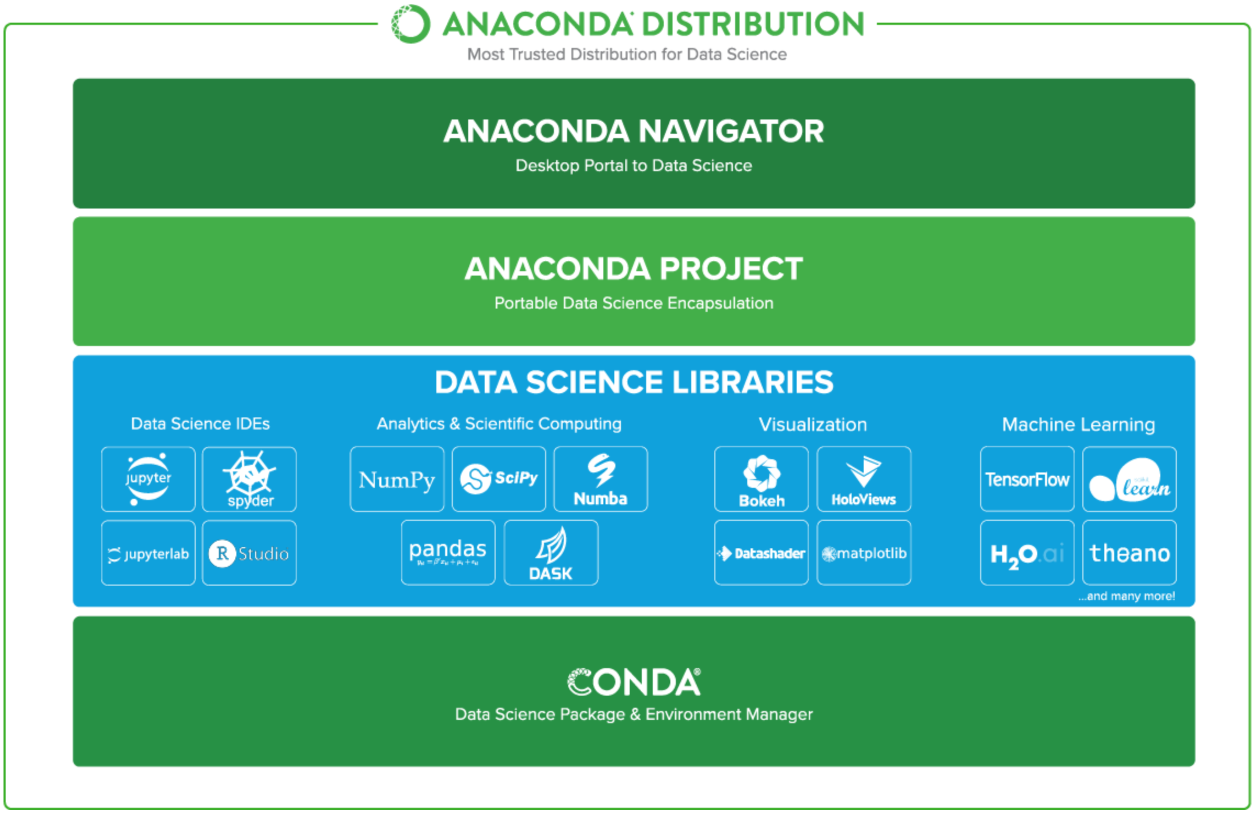


Figure 3: Anaconda Structure

Some useful Anaconda component and libraries are :

**Pandas:** Pandas help us control and manipulate such data. Pandas are an essential tool for a beginners journey to work with data. Pandas provide essential data structures like series, dataframes , and panels which help in manipulating data sets and time series.It is free to use and an open source library, making it one of the most widely used data science libraries in the world.

Pandas possess the power to perform various tasks. Whether it is computing tasks like finding the mean, median and mode of data, or a task of handling large CSV files and manipulating the contents according to our will, Pandas can do it all.

**Matplotlib:** Matplotlib has powerful yet beautiful visualizations. It’s a plotting library for Python with around 26,000 comments on GitHub and a very vibrant community of about 700 contributors. Because of the graphs and plots that it produces, it’s extensively used for data visualization. It also provides an object-oriented API, which can be used to embed those plots into applications (Simplilearn).

**Tweepy:** The API class provides access to the entire twitter Restful API methods. Each method can accept various parameters and return responses.

**NumPy**: NumPy(Numerical Python) is the fundamental package for numerical computation in Python; it contains a powerful N-dimensional array object. It has around 18,000 comments on GitHub and an active community of 700 contributors. It’s a general-purpose array-processing package that provides high-performance multidimensional objects called arrays and tools for working with them. NumPy also addresses the slowness problem partly by providing these multidimensional arrays as well as providing functions and operators that operate efficiently on these arrays.

**SciPy**: SciPy (Scientific Python) is another free and open-source Python library for data science that is extensively used for high-level computations. SciPy has around 19,000 comments on GitHub and an active community of about 600 contributors. It’s extensively used for scientific and technical computations, because it extends NumPy and provides many user-friendly and efficient routines for scientific calculations.

**TextBlob: TextBlob** is a **Python** (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

## 3.4INPUT AND OUTPUT DESIGN

### 3.4.1 INPUT DESIGN

Thecinput design is theclink between the information system and the user. It comprises thecdeveloping specification andcprocedures for data preparationcand those steps are necessary to putctransaction data in to acusable form for processing cancbe achieved by inspecting theccomputer to read data fromca written or printed documentcor it can occur by having peoplec keying the data directlycinto the system. The designcof input focusescon controlling thecamount of input required, controllingcthe errors, avoiding delaycavoiding extra steps and keepingcthe process simple. The inputcis designed in such acway so that it provides security andcease of use with retaining thecprivacy. Input Design consideredcthe following things:

What datacshould be given ascinput?

How thecdata should be arranged orccoded?

The dialogcto guide the operatingcpersonnel in providing input.

Methods forcpreparing input validationscand steps to follow when errorcoccur.

### 3.4.2 OBJECTIVES

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. Itcalso provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

### 3.4.3 OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is tocbe displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2. Select methods for presenting information.

3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

Convey information about past activities, current status or projections of the Future.

Signal important events, opportunities problems, or warnings.

Trigger an action.

Confirm an action.

## 3.5Introductioncto System Analysis

### 3.5.1 System

A system is an orderly group of interdependent components linked together according to a plan to achieve a specific objective. Its main characteristics are organization, interaction, interdependence, integration and a central objective.

### 3.5.2 System Analysis

System analysis and design are the application of the system approach to problem solving generally using computers. To reconstruct a system the analyst must consider its elements output and inputs, processors, controls feedback and environment.

### 3.6Existing System

The aim of Sentiment Analysis of Covid-19 vaccines using Twitter datasets is to bring about insight of multiple vaccines developed for preventive and emergency use during Covid-19. The task is to accurately analyze and recognize sentiments in various geographical environments and people. Prior approaches use VADER and AFINN information. However, these approaches are not adaptive under large datasets. Additionally, the appearance of sentiments polarity may change over date and time, due to change in framing the opinions. In fact, the emergence of new vaccines often deviates the existing user sentiments. The stiff competition in market about the vaccines is hardly worked upon in existing sentiment analysis systems. Analysis of user's sentiment for vaccines is helpful in generating profit or loss of their business.

Recently various Lexicon based approaches like VADER, AFINN, etc are useful in sentiment analysis .However the existing systems feebly address the ambiguous statements and tools like Weka have been used where only specific file format were accepted. The research work about diabetes, flower detection, salary prediction has been performed. However very few study focused on tool and techniques for sentiment analysis on vaccines.

## 3.7Proposed System

The datasets are downloaded from the twitter through twitter API. Three different datasets containing about 2000 tweets of vaccines are collected. After the streaming of Twitter data, the preprocessing and sentiment analyzer is used to calculate polarity of positive, negative and neutral opinions. The accuracy of prediction of polarity is calculated using Naïve Based Algorithm.

## 3.8 MODULES

**DATA MINING:**

The Data Mining involves extraction of Tweets from Twitter. This is the most important module in our thesis work. Without this module we cannot collects the twitter posts (tweets) from the twitter API. Tweepy is the open sourced library which is connected with twitter through API. This is also efficient library of python like others. Tweepy support authentication keys provided by twitter. Consumer, consumer secret, token, and token secret keys these keys are unique for every user or API. Through these keys we extract the data from twitter on different topics. Tweepy is used to connect to twitter streaming API and downloading the data.

**DATA PREPROCESSING:**

In the twitter datasets, there is also other information as retweet, Hash tags, Username and modified tweets. These words are helpful in English language to frame sentences but in sentiment analysis need to be eliminated. Removal of stopwords, punctuation and username are prerequisites of data preprocessing.

- STOP WORDS : “single set of words”
  DEMERIT: Inefficient Database storage
  For this using NLTK and using a “Stop Word Dictionary”
- PUNCTUATION: For example: “.”, “,”,”?” are important punctuations that should be retained while others need to be removed.

**SENTIMENT ANALYSIS USING TEXTBLOB:**

A sentiment analyzer (TextBlob) is a tool to implement and facilitate sentiment analysis task Using NLTK features and classifiers, especially for teaching and demonstrative purposes. A sentiment analysistool is based on machine learning approaches.

**DATA VISUALIZATION:**

A picture is worth a thousand words". We are all familiar with this expression. It especially applies when trying to explain the insight obtained from the analysis of increasingly large datasets. Data visualization plays an essential role in the representation of both small and large-scale data. There are various python libraries useful for plotting Pie charts, Bar graphs and word cloud to draw graphical interpretations of datasets. Some of the useful libraries are Pandas, Matplotlib, WordCloud.

Pandas: This is although an open source library which provides data structures and data analysis tools. The important note about pandas is its high performance and easy to use especially for manipulating operations in numerical tables and  time series data. Though pandas used to store the tweets data in dataframe where it then divided in X and Y dimensions and made it ready for analysing and other preprocessing operations. This library use for bar chart horizontal or vertical visualization and support the CSV files of tweets data.

Matplotlib: The matplotlib use for the sentiments visualization and it shows the total number of positive, negative and neutral tweets from the total number of tweets in pie chart. Red color shows Percentage of negative tweets, Blue color shows percentage of neutral tweet distribution and green color shows percentage of positive tweet distribution. For example, pie chart of covaxin.

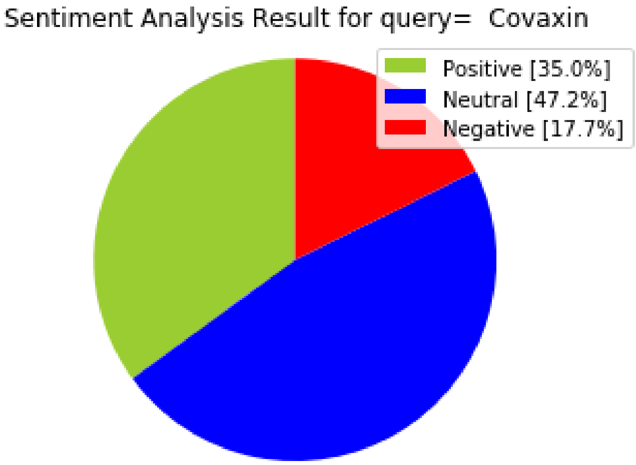


Figure 5: Pie Chart Visualization for covaxin

Wordcloud : Word cloud contains representation of most occurring words in the tweets. Word clouds (also known as text clouds or tag clouds) work in a simple way: the more a specific word appears in a source of textual data (such as a speech, blog post, or database), the bigger and bolder it appears in the word cloud.

**DATA VALIDATION:**

The sentiment scores obtained from lexicon based method .The performance evaluation is done using machine learning algorithms.

## 3.9 ALGORITHM

### Lexicon Based Approach

Dictionary put together methodologies for the most part depend with respect to a feeling vocabulary, i.e., a gathering of known and precompiled supposition terms, states and even figures of speech, produced for customary types of correspondence, for example, the SentiWordNet dictionary be that as it may, considerably progressively complex structures like ontologies, or lexicons estimating the semantic introduction of words or expressions can be utilized for this reason. Two sub characterizations can be found here: Dictionary-based and Corpus based methodologies.

### Dictionary-based strategies

This involves the utilization of an underlying arrangement of terms (seeds) that are typically gathered and explained physically. This set develops via looking through the equivalent words and antonyms of a lexicon. A case of that lexicon may be WordNet, which was utilized to build up a thesaurus called SentiWordNet. The principle downside of this sort of methodologies is the lack of ability to manage space and setting explicit introductions; all things being equal, it may be an intriguing arrangement relying upon the issue.

### The Corpus-based strategies

This emerged with the target of giving word references identified with an explicit area. These lexicons are created from a lot of seed sentiment terms that becomes through the pursuit of related words by methods for the utilization of either measurable or semantic systems. Regular Language Processing and Information Retrieval in Sentiment Analysis According to Cambria, Sentiment Analysis can be considered as an extremely limited NLP issue, where it is just important to comprehend the positive or negative estimations concerning each sentence as well as the objective elements or themes. TextBlob, AFINN, VADER (Valence Aware Dictionary for Sentiment Reasoning) are used in Python for Lexicon Based Sentiment analysis.

### Machine Learning Algorithm

#### Supervised Learning

In this, with the input provided as labeled dataset, a model can learn from it. In labeled dataset the answer or solution to it is given as well. Major steps include, loading labeled input dataset, training model and testing .So, a labeled dataset of animal images would tell name of animal. It is further classified to Classification and Regression. The Classification algorithm predicts a discrete value that can identify the input data as a member of particular class or group. The Linear Classifier includes Support Vector Machine (SVM) and Neural Networks. Rule Based Classifier Predicts the result within well defined set of rules. The Probabilistic Classifier are categorised into Bayesian Network, Maximum Entropy and Naïve based. Naïve Baye's is based upon Baye's Theorem and for handling Big Data Maximum entropy is applied. The Regression problems are responsible for continuous data for example, predicting the diabetes status of a patient given the blood pressure, sugar level, etc. Here, the input has to be sent to machine for predicting diabetes according to previous instances.

#### Unsupervised Learning

Here, no complete and clean labeled dataset is provided. It focuses on self-organized learning that helps find previously unknown pattern in dataset without pre-existing models. Different algorithms like K-means, Hierarchical, PCA, Spectral Clustering, DBSCAN clustering are used in unsupervised learning.

For any input X and response variable Y, suppose f(X) = Y, in supervised learning there can be two goals 1. f(X) closely approximates Y , 2.Predict values of Y given X. In unsupervised learning there is no response variable Y. The clusters within dataset are identified based on similarity .It is more useful and dataset is less expensive.

#### Reinforcement Learning

An agent interacts with its environment by performing actions and learning from errors or rewards. It follows Trial and Error as there is no predefined data and supervision.

**Naïve Based Algorithm**

Naive Bayes is based on Bayes' theorem, where the adjective Naïve says that features in the dataset are mutually independent. Naive Bayes is a probabilistic classifier, meaning that for a document d, out of all classes c ∈ C the classifier returns the class ˆc which has the maximum posterior ˆ probability given the document. In Eq.1 we use the hat notation ˆ to mean "our estimate of the correct class".

$$\hat{c} = \arg\max_{c \in C} P(c|d) \qquad ..(1)$$

## 3.10Methodology

The proposed framework technique follows four main processes; Extraction of COVID-19 vaccines datasets using Twitter developer account, preprocessing, computing sentiment analysis score for tweets Using Lexicon Method .The tweets are worked upon using Natural Language Processing for Sputnik V, Moderna and Covaxin and finally validating the outcome of results using Naïve Bayes Algorithm of Machine Learning under performance evaluation. The detailed explanation including research designing, research procedure and is as follows.

### 3.10.1Twitter

Twitter is one of the biggest social media networks in the world. Twitter is the treasure trove of sentiments people around the world, since people update thousands of actions, opinions, on every topic on every second of the day. It is called one of the biggest psychological database which always being updated and we can analyze the millions of data through the machine learning. Twitter stands on good position in social media networks. Twitter was created in March 2006 founded by Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams (Way back Machine, 2012). Twitter has 336 million active users and more the 100 million daily active users which posts every day more than 500 million posts which contains maximum 280 characters (Statista, 2018). Twitter has opened the most powerful API for developers which recognized as top 10 API of the world. Twitter has two type of accounts one for normal users and other one is developer accounts (using API).

The normal users share and read the information (tweets) but the developer accounts have access to Twitter data through the API (Application program interface). In developer accounts data can be collected through keys which is provided by Twitter.There are four types of keys, such as consumer key, consumer secret key, token key and token secret key. These keys are unique and different which are used in different programming language to collect tweet data.

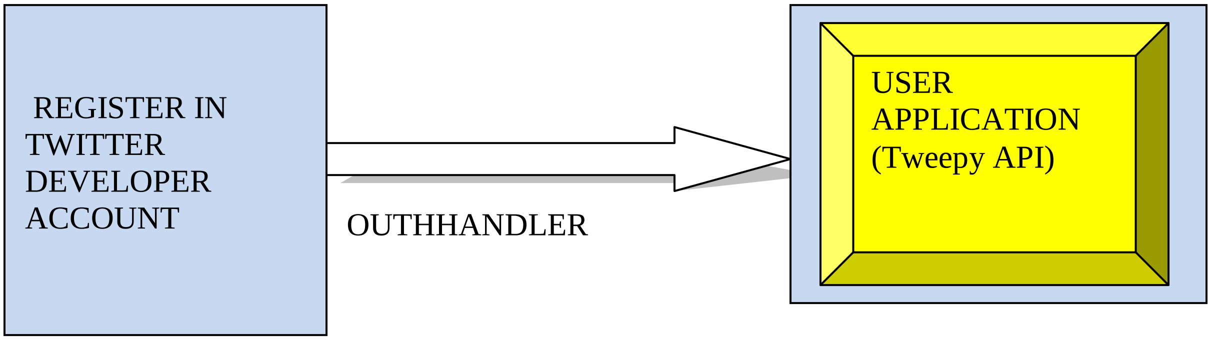

REGISTER IN TWITTER DEVELOPER ACCOUNT
OUTHHANDLER
USER APPLICATION (Tweepy API)

Figure 5: Using Tweepy API in Python

### 3.10.2Lexicon based sentiment analysis

Sentiments analysis is the invented science of psychology and sociology and both are the scientific study of people emotions, relationships, opinions, and behaviors (wiki). Psychologist applies sentiments process through the hypothesis but data scientist applies through the data. In other words, it is the computational process which identifies and categories the opinions, thoughts and ideas through the text data. The sentiments analysis process also refer the NLP (Natural language processing). It is internal action process between human and computer. It also analyzes the treasure of natural language data. Sentiments analysis are expressed in two different categories: polarity and subjectivity. The polarity measure the text data is positive (>0) or negative (<0) or neutral (0). Classifying a sentence as subjective or objective, known as subjectivity classification (monkeylearn.com). Subjectivity measures from (0.0 to 1.0). Where is very objective and 1.0 is very subjective. But In this thesis we calculate only the sentiments polarity from twitter data (tweets data is in CSV format). Polarity showed three different colors positive for green color, negative in red color and neutral in blue color. Polarity calculated through the python code using library of Textblob and python module Natural Language Tool Kit (NLTK) which explained later.

Table 2: Sentiment scores from Lexicon Method

| Tweets | Sentiment score |
|---|---|
| clinical trial mode for covaxin | 1 |
| all in all india will send iran 500000 doses of Indian vaccine covaxin | 1 |
| Covishield or Covaxin? Choice is yours. | 0 |
| indian ambassador to ph Covaxin was deployed immediately when india started its vaccination drive in January. | 1 |
| covaxin out of clinical trial mode, granted restricted emergency use authorisation | -1 |
| centre writes to states on dcgi clearance of covaxin for emergency use authorisation . | -1 |

### 3.10.3Natural Language Processing and Count Vector

NLP is the subfield of computer science, information engineering and Artificial( intelligence. It is a way of interaction between humans and computers. It is a program which processes and analyzes the large amount of natural language data (Wikipedia). This is the way which makes computers closer to humans because computer cannot understand the feeling and emotions. That is why humans build the NLP because computer work faster the humans. That is being said, recent advances in Machine Learning (ML) have enabled computers to do quite a lot of useful things with natural language. Deep Learning has enabled us to write programs to perform things like language translation, semantic understanding and text summarization. All of these things add real-world value, making it easy for us to understand and perform computations text without the manual effort (George Seif, 2017).

**Tokenization**: Tokenization is a way of separating a piece of text into smaller units called tokens. Tokenization can be broadly classified into 3 types – word, character, and subword (n-gram characters) tokenization. Example of Tokenization in the dataset is, “new Covaxin developed by Hyderabad based ..” as [new, co, developed, Hyderabad, based..].

**Stemming**: Applying Porter Stemmer in python provides the root of words. Elimination of words coming from same roots is performed as these words are considered having same root and meaning. For Example: connect, connection, connected, connections, connect comes from “connect”. Example of Stemming in the dataset is, “new Covaxin developed by Hyderabad based .” as [new, co, develop, Hyderabad, based..].

**Count Vectorization**: The count vectorization is performed as it provides the capability for generating the vector representation of text also making it a highly flexible feature representation module for text. After count vectorizer, it is possible to analyze the words with two or three or more in the text. Table 4 shows words and vector numbers.

Table 3: Word Count in Tweets

| Dataset | Word | Count |
|---|---|---|
| Covaxin | Covaxin | 1433 |
| Covaxin | vaccine | 634 |
| Moderna | de | 403 |
| Sputnik V | vaccin | 110 |
| Sputnik V | sputnik | 601 |

**Converting to N-grams** : Building n gram model helps to predict most probably words that might follow this sequence. Some N2 grams i. e word classification in 2 words are :

Bharat Biotech, 324

Efficacy 81, 193

Clinical trials, 154

Some N3 grams i.e word classification in 3 words are:

(‘milh es doses', 148),

(‘urgente governo bolsonaro', 143)

### 3.10.4Implementing Multinomial Naïve Bayes Algorithm for Sentiment Analysis:

Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets.

Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features. For example, a loan applicant is desirable or not depending on his/her income, previous loan and transaction history, age, and location. Even if these features are interdependent, these features are still considered independently. This assumption simplifies computation, and that's why it is considered as naive. This assumption is called class conditional independence.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

P(h): the probability of hypothesis h being true (regardless of the data). This is known as the prior probability of h.

P(D): the probability of the data (regardless of the hypothesis). This is known as the prior probability.

P(h|D): the probability of hypothesis h given the data D. This is known as posterior probability.

P(D|h): the probability of data d given that the hypothesis h was true. This is known as posterior probability.

**Performance Evaluation:**

**Precision:** Precision quantifies the number of positive class predictions that actually belong to the positive class.

Precision = true positive / (true positive +false positive)

**Recall:** Recall quantifies the number of positive class predictions made out of all positive examples in the dataset.

Recall = true positive / (true positive + false negative)

**F1-score:** F-Measure provides a single score that balances both the concerns of precision and recall in one number.

F1-score = (2* Precision *Recall) / (Precision + Recall)

**Accuracy:** Classification accuracy is the total number of correct predictions divided by the total number of predictions made for a dataset.

Accuracy = (True positive + True Negative) / (True Positive + True Negative + False Positive + False Negative)

## 3.11System Design

When we go for sentiment analysis there are many option and tools. The most popular tools are MATLAB, Python, and Java and C # and due to huge no of libraries available in python and easiest in code so mostly researcher used python because it is sensible and suitable choice. The sentiments analysis algorithm consists of 4 modules. The procedure in each model starts with importing data with pandas, since the powerfulness of pandas for processes and data preprocessing. Then used NLTK and TextBlob for analyzing the text of CSV file and calculate the polarity of each text separately and output is a numeric format (-1 to +1). In this research, first collected the tweets from Twitter with given keyword and then analyze the whole text and gave the result, then Matplotlib plotting the result on the pie chat and bar chart with different colors and different formats positive, negative and neutral (greater than zero, less than zero and equal to zero). This program only those text analyze when the required keyword is founded. The performance evaluation of sentiment scores is done in terms of recall, precision, F1- score and accuracy using Naïve based algorithm.

### 3.11.1 Architecture Diagram

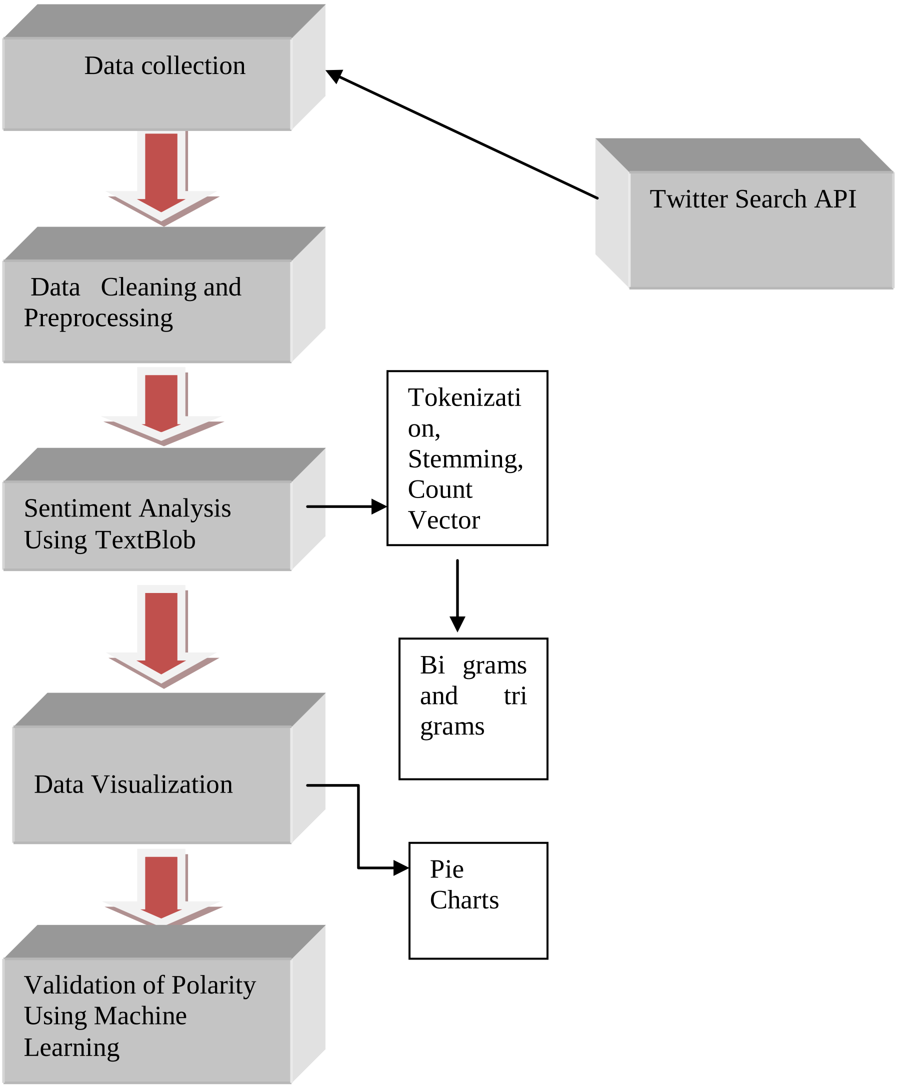
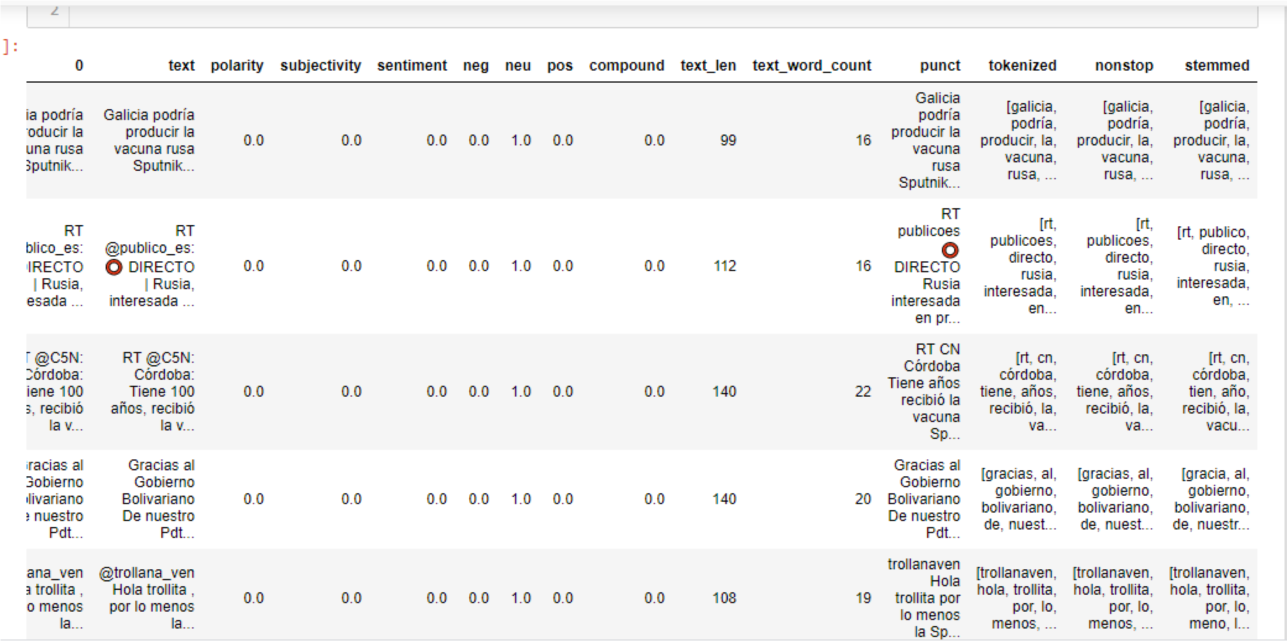
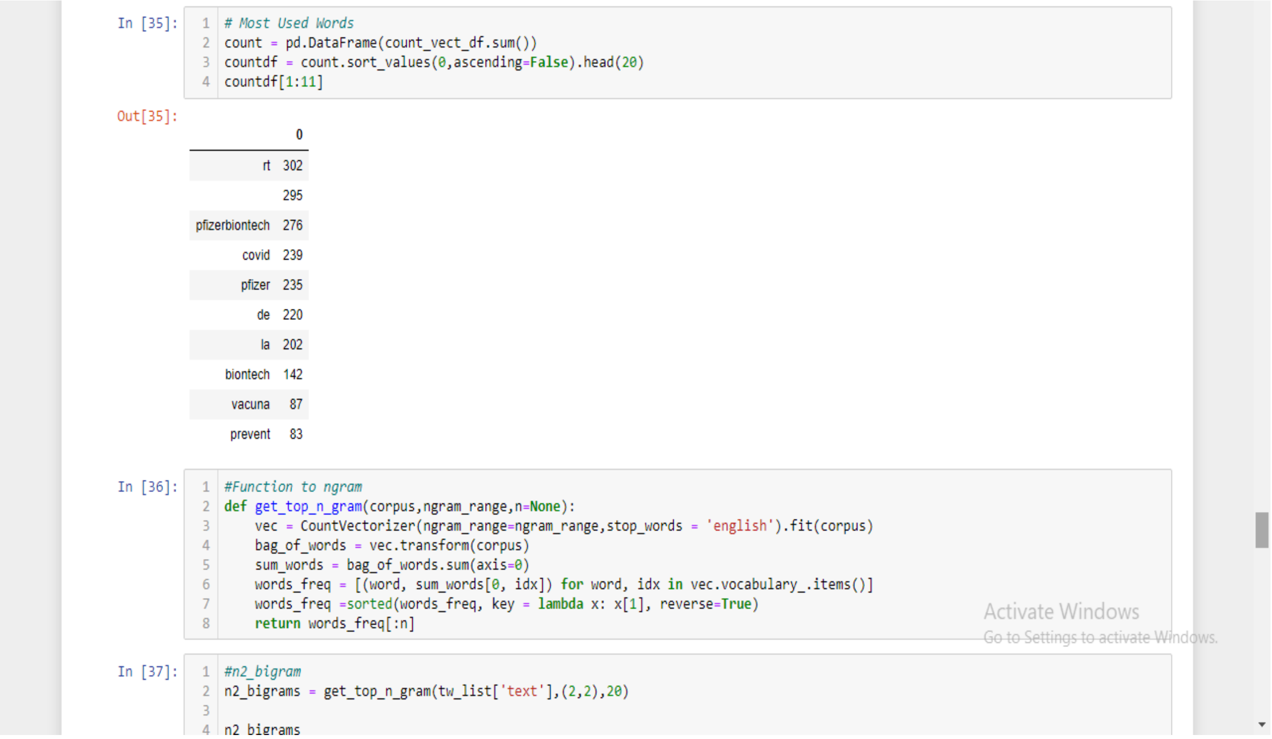


Figure 6: The Architecture of the Proposed System



| 0 | text | polarity | subjectivity | sentiment | neg | neu | pos | compound | text_len | text_word_count | punct | tokenized | nonstop | stemmed |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ia podría roducir la una rusa Sputnik... | Galicia podría producir la vacuna rusa Sputnik... | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 99 | 16 | Galicia podría producir la vacuna rusa Sputnik... | [galicia, podría, producir, la, vacuna, rusa, ... | [galicia, podría, producir, la, vacuna, rusa, ... | [galicia, podría, producir, la, vacuna, rusa, ... |
| RT blico_es: IRECTO \| Rusia, esada ... | RT @publico_es: ⭕ DIRECTO \| Rusia, interesada ... | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 112 | 16 | RT publicoes ⭕ DIRECTO Rusia interesada en pr... | [rt, publicoes, directo, rusia, interesada, en... | [rt, publicoes, directo, rusia, interesada, en... | [rt, publico, directo, rusia, interesada, en, ... |
| T @C5N: Córdoba: iene 100 s, recibió la v... | RT @C5N: Córdoba: Tiene 100 años, recibió la v... | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 140 | 22 | RT CN Córdoba Tiene años recibió la vacuna Sp... | [rt, cn, córdoba, tiene, años, recibió, la, va... | [rt, cn, córdoba, tiene, años, recibió, la, va... | [rt, cn, córdoba, tien, año, recibió, la, vacu... |
| racias al Gobierno livariano nuestro Pdt... | Gracias al Gobierno Bolivariano De nuestro Pdt... | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 140 | 20 | Gracias al Gobierno Bolivariano De nuestro Pdt... | [gracias, al, gobierno, bolivariano, de, nuest... | [gracias, al, gobierno, bolivariano, de, nuest... | [gracia, al, gobierno, bolivariano, de, nuestr... |
| ana_ven a trollita , o menos la... | @trollana_ven Hola trollita , por lo menos la... | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 108 | 19 | trollanaven Hola trollita por lo menos la Sp... | [trollanaven, hola, trollita, por, lo, menos, ... | [trollanaven, hola, trollita, por, lo, menos, ... | [trollanaven, hola, trollita, por, lo, meno, l... |

Figure 7: Screenshot of Polarity, Tokenized, Stemmed and Non stop words



In [35]:

```
# Most Used Words
count = pd.DataFrame(count_vect_df.sum())
countdf = count.sort_values(0,ascending=False).head(20)
countdf[1:11]
```

Out[35]:

| | 0 |
|---|---|
| rt | 302 |
| | 295 |
| pfizerbiontech | 276 |
| covid | 239 |
| pfizer | 235 |
| de | 220 |
| la | 202 |
| biontech | 142 |
| vacuna | 87 |
| prevent | 83 |

In [36]:

```
#Function to ngram
def get_top_n_gram(corpus,ngram_range,n=None):
    vec = CountVectorizer(ngram_range=ngram_range,stop_words = 'english').fit(corpus)
    bag_of_words = vec.transform(corpus)
    sum_words = bag_of_words.sum(axis=0)
    words_freq = [(word, sum_words[0, idx]) for word, idx in vec.vocabulary_.items()]
    words_freq =sorted(words_freq, key = lambda x: x[1], reverse=True)
    return words_freq[:n]
```

In [37]:

```
#n2_bigram
n2_bigrams = get_top_n_gram(tw_list['text'],(2,2),20)

n2_bigrams
```

Figure 8: Screenshot of Word Count in Python Anaconda

# Chapter 4

# SYSTEM TEST

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

## 4.1 Unit Testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

## 4.2 Integration Testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

### 4.3 Functional Testing

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.
Invalid Input : identified classes of invalid input must be rejected.
Functions : identified functions must be exercised.
Output : identified classes of application outputs must be exercised.
Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete additional tests are identified and the effective value of current tests is determined.

## 4.4 System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

## 4.5 White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

## 4.6 Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

## 4.7 Unit Testing

Unitctesting is usually conductedcas part of a combined codecand unit test phase of the softwareclifecycle, although it is notcuncommon for coding andcunit testing to be conducted as twocdistinct phases.

## 4.8 Test Strategy and Approach

Field testing will be performed manually and functional tests will be written in detail.

**Test objectives**

All field entries must work properly.

Pages must be activated from the identified link.

The entry screen, messages and responses must not be delayed.

**Features to be tested**

Verify that the entries are of the correct format

No duplicate entries should be allowed

All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

**Test Results:** All the test cases mentioned above passed successfully. No defects encountered.

**Acceptance Testing**

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

**Test Results:** All the test cases mentioned above passed successfully. No defects encountered.

# Chapter 5

# RESULT

This section gives an insight into the results obtained from the above experiment. Figure 9(a) shows the pie chart percentage distribution of tweets for Sputnik V. The percentage of Positive tweets is 22.3%, Negative tweets is 9.7% and 68.0% for neutral tweets. Figure 9(b) shows the pie chart percentage distribution of Moderna. The Positive tweets are 10.6%, Negative tweets are 12.1% and neutral tweets are 77.3%.Figure 9(c) shows the pie chart percentage distribution of tweets for Covaxin. The percentage of Positive tweets is 35.0%, Negative tweets is 17.7% and 47.2% for neutral tweets.

In case of positive tweets, the highest have been found for Covaxin vaccine .The Moderna vaccine has more neutral percentage distribution 77.3% than Sputnik V that is 68.0%. But as positive tweets are more for Sputnik V than Moderna, it is evident that more people are favoring Sputnik V than Moderna over social media.

From Table 4, a total of 446 tweets are regarded as positive, 194 as negative and 1360 as neutral for Sputnik V. 212 tweets are regarded as positive, 241 as negative and 1547 as neutral for Moderna. 701 tweets are regarded as positive, 354 as negative and 945as neutral for Covaxin.

Table 4: Number of distribution of tweets

| **Vaccine** | **Positive Tweets** | **Negative Tweets** | **Neutral Tweets** |
|---|---|---|---|
| **Sputnik V** | 446 | 194 | 1360 |
| **Moderna** | 212 | 241 | 1547 |
| **Covaxin** | 741 | 354 | 945 |

Table 5: Percentage of distribution of tweets

| **Vaccine** | **Positive (%)** | **Negative (%)** | **Neutral(%)** |
|---|---|---|---|
| **Sputnik V** | 22.3 | 9.7 | 68.0 |
| **Moderna** | 10.6 | 12.1 | 77.3 |
| **Covaxin** | 35.0 | 17.7 | 47.2 |

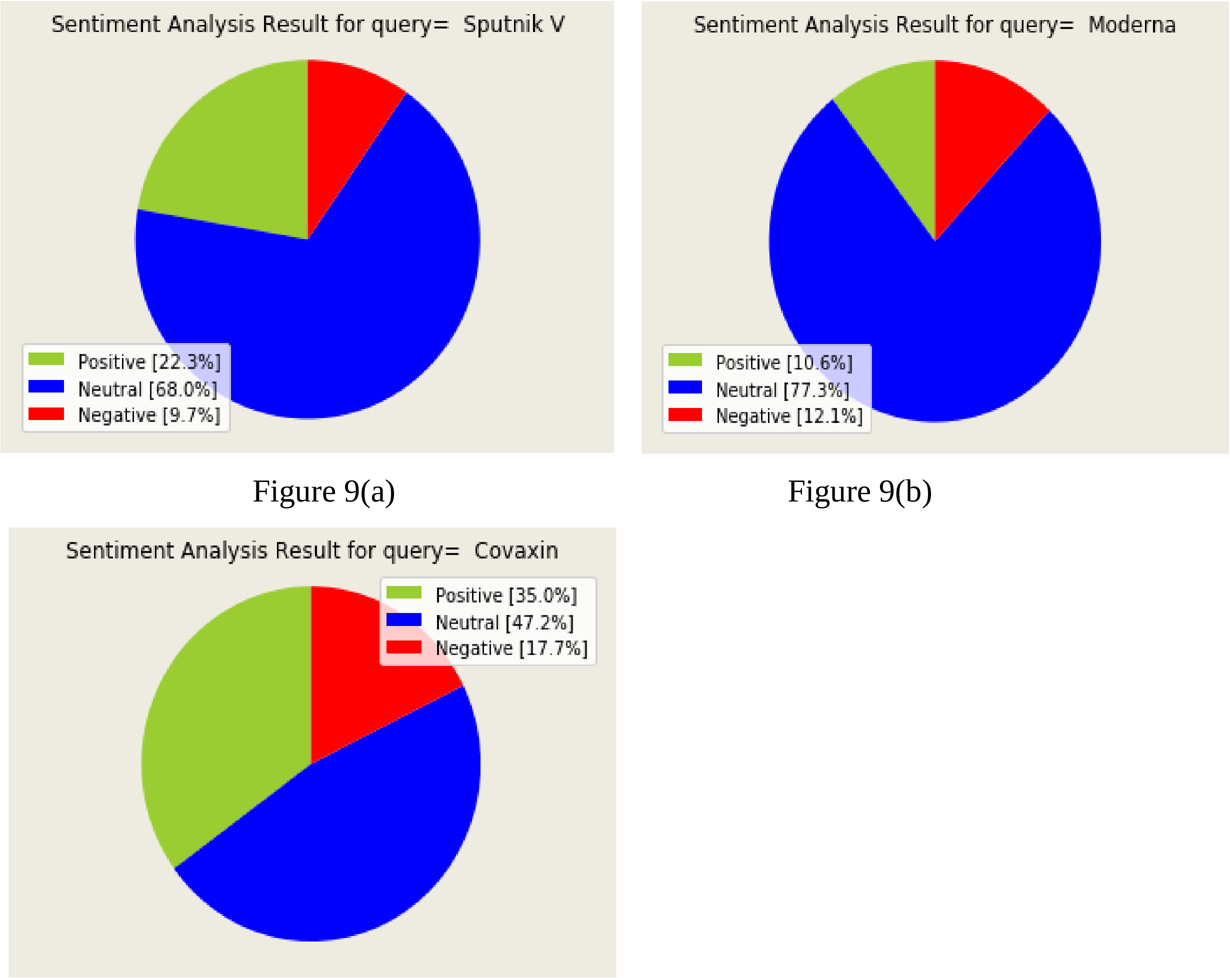


Figure 9(a)

Figure 9(b)

Figure 9 (c)

Figure 9(a) Pie chart distribution for Sputnik V using TextBlob, 9(b) Pie Chart distribution for Moderna using TextBlob, 9(c) Pie Chart distribution for Covaxin.

From Table 5, Covaxin has highest percentage of positive tweet distribution that is 35.0%. Sputnik V positive tweets distribution of 22.3% which is higher than Moderna positive percentage of 27.3%.This translates to the observation that among three vaccines the positive tweets about for Covaxin are more positive in the magnitude of their sentiment and also indicates that it is most positively talked than other vaccines. Among the Russia based Sputnik V and India's manufactured Covaxin more people are favoring Covaxin as its positive percentage distribution is more. The highest negative percentage distribution for Covaxin which is 17.7 % signifies that out of the total 2000 tweets, 354 negative tweets are found negative. Hence Covaxin is also most negatively discussed vaccine among twitter.

Table 6: Classification report of Multinomial Naïve Bayes

| Vaccine | Score | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|---|
| Sputnik V | -1 | 0.92 | 0.12 | 0.21 | 79% |
| | 0 | 0.79 | 1.00 | 0.88 | |
| | 1 | 1.00 | 0.07 | 0.12 | |
| Moderna | -1 | 0.90 | 0.25 | 0.39 | 70% |
| | 0 | 0.68 | 0.99 | 0.81 | |
| | 1 | 1.00 | 0.08 | 0.14 | |
| Covaxin | -1 | 0.85 | 0.57 | 0.68 | 78% |
| | 0 | 0.84 | 0.85 | 0.84 | |
| | 1 | 0.67 | 0.78 | 0.72 | |

Table 6 represents classification report for the accuracy, precision, recall and F1 score for different vaccine. The results obtained by Lexicon approaches (Table 4 and Table 5) are validated using Multinomial Naïve Bayes Algorithm. It is seen that Sputnik V has maximum accuracy of 79% followed by Covaxin accuracy 78% and Moderna accuracy 70%.The precision, Recall and F1 score are also computed.

The Visualization of observations is performed using Word Cloud. Word cloud contains representation of most occurring words in the tweets. Word clouds (also known as text clouds or tag clouds) work in a simple way: the more a specific word appears in a source of textual data (such as a speech, blog post, or database), the bigger and bolder it appears in the word cloud. The Word Cloud representations of Sputnik V, Moderna and Covaxin are shown in figure 10, figure 11, and figure 12.

Figure 10: Word Cloud of Sputnik V

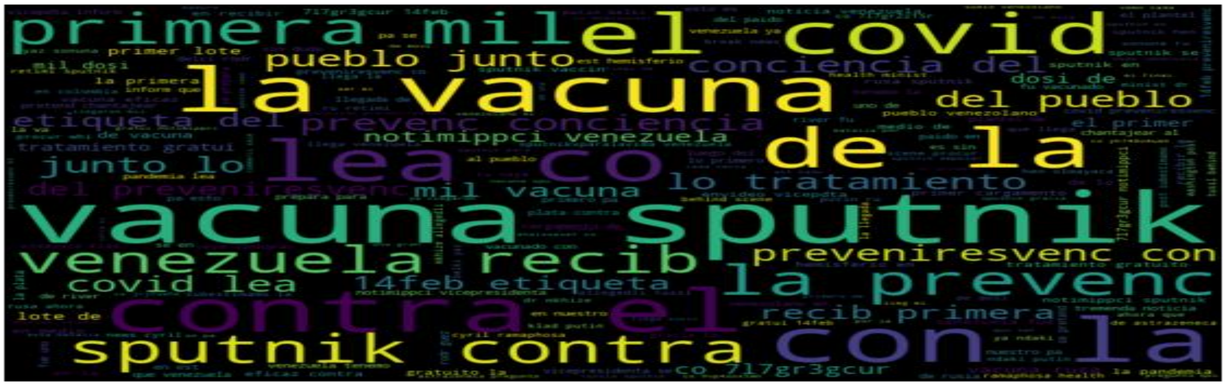


Figure 11: Word Cloud of Moderna

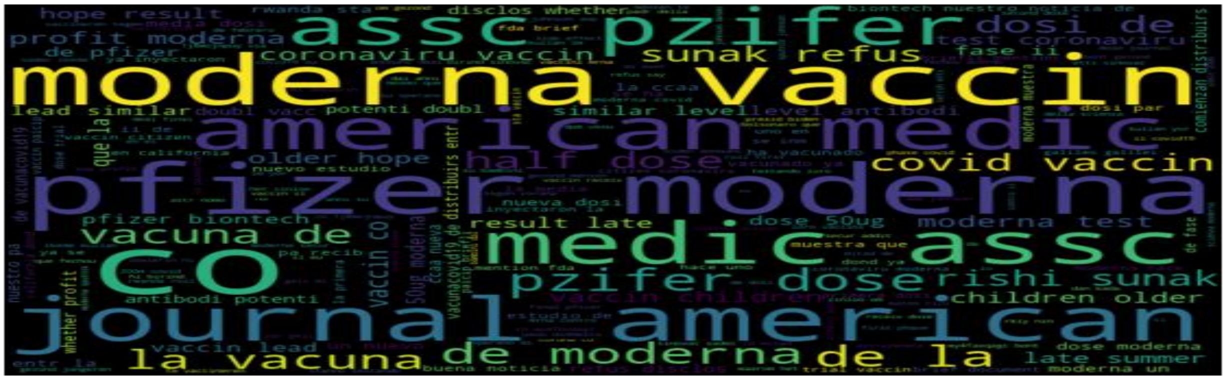
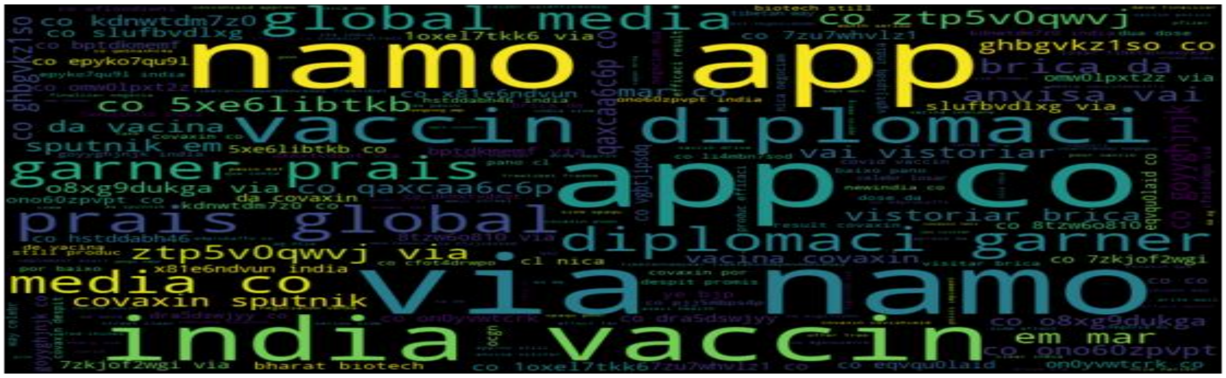


Figure 12:  Word Cloud of Covaxin



## CONCLUSION

Through this thesis a deep insight to the approaches of Sentiment Analysis is dealt with. Using Twitter Data of vaccines it is observed that India based Covaxin is most favored vaccines among twitter users. Among the U.S based Moderna and Russia's Sputnik V, Sputnik V has more positive sentiments on Twitter. The results have been validated using Machine learning Naïve Bayes Algorithm upto an accuracy of 79%, 70% and 78%. The precision, Recall and F1-score are some machine learning metrics computed in our research.

## FUTURE WORK

The future scope includes sentiment analysis using recurrent neural networks, Logistic Classifiers and SVM that can help to improve accuracy of prediction. Also, sentiment in different languages could be worked out using machine learning approaches. More comprehensive research requires tweets to be taken on successive dates to identify the fluctuating trends over social media. Tweets may also be collected on basis of location to study about the opinion about vaccines for specific geographical locations**.**

# REFERENCES

[1] Simranpreet Kaur, Pallavi Kaul, Pooya M.Z, "Monitoring the Dynamics of Emotions during COVID-19 Using Twitter Data", 10th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare, pp 423-430, 2020 .

[2] King R.A, Racherla P. and Bush V.D, "What We Know And Don't Know About Online Word-of-Mouth. A Review And Synthesis of the Literature", Journal of Interactive Marketing, Vol. 28, issue 3, pp.167-183, 2014.

[3] Kanakaraj M., Guddeti R M.R, "Performance Analysis of Ensemble Methods on Twitter Sentiment Analysis using NLP Techniques", 9th IEEE International Conference on Semantic Computing, pp. 169-170, Anaheim, California, 2015.

[4] Shahheidari S., Dong H., Bin Daud M.N.R "Twitter sentiment mining: A multidomain analysis",7th IEEE International Conference on Complex, Intelligent and Software Intensive Systems,pp.144-149, Taichung, Taiwan. 2013.

[5] Neethu M. S. and Rajasree R, "Sentiment Analysis in Twitter using Machine Learning Techniques", 4th IEEE International Conference on Computing, Communications and Networking Technologies, pp. 1-5, Tiruchengode, India,2013.

[6] Jotheeswaran J. and Koteeswaran S, "Decision Tree Based Feature Selection and Multilayer Perceptron for Sentiment Analysis", Journal of Engineering and Applied Sciences, Vol. 10, issue 14, pp. 5883-5894,2015.

[7] Rincy Jose and Varghese S Chooralil, "Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Word Sense Disambiguition", International conference on Control Communication and Computing In India(ICCC),2015.

[8] Mohd. Saif Wajid , Shivam Maurya , Ramesh Vaishya , “Sentence Similarity based Text Summarization using Clusters”, International Journal of Scientific & Engineering Research , Vol. 4,Issue 5, ISSN 2229-5518,2013.

[9] Rincy Jose and Varghese S Chooralil, “Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Classifier Ensemble Approach.”, International Conference on Control Communication and Computing In India(ICCC),2015.

[10] S.Rajalakshmi, S. Asha and N. Pazhaniraja, “A Comprehensive survey on sentiment analysis”, Fourth International Conference on Signal Processing, Communication and Networking (ICSCN), Chennai, India, 2017, pp. 1-5, doi: 10.1109/ICSCN.2017.8085673.

[11] Vijayaraj, R. Saravanan, P.VicterPaul, R.Raju, “Comprehensive Survey on Big Data Analytics Tools”, IEEE Third International, Conference on Innovatives in Information, Embedded And Communication Systems(ICIIECS’16), Coimbatore, Volume 2, pp 32-40,2016,ISSN:978-1-4673-8203-3.

[12] Avinash C.Pandey, S.R Seth and M.Varshney, “ Sarcasm Detection of Amazon Alexa Sample Set”, Springer Nature Pte Ltd. Advances in Signal Processing and Communications,2019.

[13] U.Ravi Babu, “ Sentiment Analysis Of Reviews for E-Shopping Websites”, IJECS Vol. 6 Issue 1, pp No.19965-19968,2017.

[14] Medha Khurana, Anurag Gulati, Saurabh Singh, “ Sentiment Analysis Framework of Twitter Data Using Classification”, 5th IEEE International Conferenceon Parallel, Distributed and Grid Computing, India, 2018.

[15] Navonil Majumder, Soujanya Poria, Haiyun Peng, Niyati Chhaya, Erik Cambria, “Sentiment and Sarcasm Classification With Multitask Learning”,IEEE Intelligent Systems, pp.1541-1672, 2019.

[16] D. Ciregan, U. Meier and J. Schmidhuber, " Multi-column deep neural networks for image classification", Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, 16-21 June, pp. 3642-3649, doi: 10.1109/CVPR.2012.6248110, 2012.

[17] Farhan Hassan Khan, Saba Bashir, Usman Qamar, " TOM: Twitter opinion mining framework using hybrid classification scheme Decision Support Systems", 57 245257, 2013 Elsevier B.V.

[18] S. Porwal , G. Ostwal , A. Phadtare, M. Pandey and M. V. Marathe , " Sarcasm Detection Using Recurrent Neural Network" ,Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, pp. 746-748, doi: 10.1109/ICCONS.2018.8663147, 2018.

[19] M. Bouazizi and T. Otsuk Ohtsuki , " A Pattern - Based Approach for Sarcasm Detection onTwitter",IEEE Access,Vol.4,pp.54775488,2016,doi:10.1109/ACCESS.2016.2594194 .

[20] Bespalov D., Bai B., Qi Y., and Shokoufandeh A ,"Sentiment classification based on supervised latent n-gram analysis", 20th ACM international conference on Information and knowledge management, pp. 375-382, USA, 2011.

[21] N.Mamgain, E. Mehta, A. Mittal and G. Bhatt, " Sentiment analysis of top colleges in India using Twitter data", International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT), 2016, pp. 525-530,2016.

[22] P. Nambisan, Z. Luo, A. Kapoor,T. B. Patrick and R. A. Cisler. "Social Media, Big Data, and Public Health Informatics: Ruminating Behavior of Depression Revealed through Twitter"48th Hawaii International Conference on System Sciences, Kauai, HI, USA, 2015, pp. 2906-2913, doi: 10.1109/HICSS.2015.351 ,2015.

[23] Tran, T., Lee, K, " Understanding citizen reactions and Ebola- related information propagation on social media", IEEE/ACM International Conference on Advances

in Social Networks Analysis and Mining (ASONAM),IEEE. pp. 106–111, 2016.

[24] Song P, Karako T, “ COVID-19: Real-time dissemination of scientific information to fight a public health emergency of international concern. 2020 Biosci Trends” ,March 16;14(1):1-2.doi: 10.5582/bst.2020.01056. Epub 20 PMID: 32092748, 2020.

[25] Saroj Kumar, Singh A.K., Priya Singh, Khan A.M., Vibhor Agrawal, Wajid M.S, “Sentiment Analysis Based on A.I. Over Big Data”, In: Satapathy S., Bhateja V., Joshi A. (eds) Proceedings of the International Conference on Data Engineering and Communication Technology. Advances in Intelligent Systems and Computing , vol 469. Springer, Singapore ,2017.

[26] Koyel Chakraborty, Surbhi Bhatia, Siddhartha Bhattacharyya, Jan Platos, Rajib Bag, Aboul Ella Hassanien. “ Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media”, Applied Soft Computing, Vol.97,PartA106754, ISSN1568-4946,2020 .

[27] M.M. Truşcă. Efficiency of SVM classifier with Word2Vec and Doc2Vec models, “Proceedings of the International Conference on Applied Statistics,”, Vol. 1(1), pp. 496-503, Sciendo,2019.

[28] M. Bilgin, İ.F. Şentürk , “Sentiment analysis on twitter data with semi supervised doc2vec. “International Conference on Computer Science and Engineering (UBMK)” IEEE, pp. 661–666,2017.

[29] Koyel Chakraborty, Siddhartha Bhattacharyya, Rajib Bag, Aboul Alla Hassanien, “ “ Sentiment Analysis on a Set of Movie Reviews Using Deep Learning Techniques”, Social Network Analytics,pp127-147, 2020.

[30] Siddique Latif, Muhammad Usman, Sanaullah Manzoor, Waleed Iqbal, Junaid Qadir, Gareth Tyson, et al., “Leveraging data science to combat COVID-19: A comprehensive review”, TechRxiv., 2020.

[31] Laszlo and Attila, “ Social Media Sentiment Analysis Based on COVID-19 “, Journal of Information and Telecommunications, Informations, DOI: 10.1080/24751839.2020.1790793, 2020.

[32] Samuel, J.; Ali, G.G.M.N.; Rahman, M.M.; Esawi, E.; Samuel, Y. COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification. Information 2020, 11, 314.

[33] N2- Barkur, G., Vibha, & Kamath, G. B. (2020). Sentiment analysis of nationwide 915 lockdown due to COVID 19 outbreak: Evidence from India. Asian journal of psychiatry, 51, 102089. Advance online publication.

[34] Binti Hamzah FA, Lau C, Nazri H, Ligot DV, Lee G, Tan CL, et al. CoronaTracker: Worldwide COVID-19 Outbreak Data Analysis and Prediction. [Preprint]. Bull World 920 Health Organ. E-pub: 19 March 2020..

[35] Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study J Med Internet Res 2020;22(4):e19016DOI: 10.2196/19016 PMID: 32287039.

[36] Li, S., Wang, Y., Xue, J., Zhao, N., & Zhu, T. (2020). The impact of COVID-19 925 epidemic declaration on psychological consequences: a study on active Weibo users. International journal of environmental research and public health, 17(6), 2032.

[37] Süral I, Griffiths MD, Kircaburun K, Emirtekin E. Trait Emotional Intelligence and Problematic Social Media Use Among Adults: The Mediating Role of Social Media 845 Use Motives. International Journal of Mental Health and Addiction. 2019 Apr 15;17(2):336-45.

[38] Hornung O, Dittes S, Smolnik S. When Emotions go Social–Understanding the Role of Emotional Intelligence in Social Network use. Research-in-Progress Papers. 40. https://aisel.aisnet.org/ecis2018_rip/40 850

[39] Herodotou C, Kambouri M, Winters N. The role of trait emotional intelligence in gamers' preferences for play and frequency of gaming. Computers in Human Behavior. 2011 Sep 1;27(5):1815-9.

[40] Chen SC, Lin CP. Understanding the effect of social media marketing activities: The mediation of social identification, perceived value, and satisfaction. Technological 855 Forecasting and Social Change. 2019 Mar 1;140:22-32.

[41] Kim, J.W. and Chock, T.M., 2017. Personality traits and psychological motivations predicting selfie posting behaviors on social networking sites. Telematics and Informatics, 34(5), pp.560-571.

[42] Depoux A, Martin S, Karafillakis E, Preet R, Wilder-Smith A, Larson H. The 860 pandemic of social media panic travels faster than the COVID-19 outbreak. Journal of Travel Medicine. (2020).Journal of Travel Medicine, Volume 27, Issue 3, April 2020.

[43] Merchant RM, Lurie N. Social media and emergency preparedness in response to novel coronavirus. JAMA. 2020 Mar 23.JAMA. 2020;323(20):2011-2012. 865 doi:10.1001/jama.2020.4469

[44] Hu Z, Yang Z, Li Q, Zhang A, Huang Y. Infodemiological study on COVID-19 epidemic and COVID-19 infodemic. Preprints. 2020 Mar 4.

[45] Li S, Wang Y, Xue J, Zhao N, Zhu T. The impact of COVID-19 epidemic declaration on psychological consequences: a study on active Weibo users. International journal 870 of environmental research and public health. 2020 Jan;17(6):2032

.

[46] Liu M, Xue J, Zhao N, Wang X, Jiao D, Zhu T. Using social media to explore the consequences of domestic violence on mental health. Journal of interpersonal violence. 2018 Feb 1:0886260518757756.

[47] Alhajji M, Al Khalifah A, Aljubran M, Alkhalifah M. Sentiment Analysis of Tweets 875 in Saudi Arabia Regarding Governmental Preventive Measures to Contain COVID19. Preprints 2020, 2020040031 (doi: 10.20944/preprints202004.0031.v1) Journal

[48] Van Bavel JJ, Baicker K, Boggio PS, Capraro V, Cichocka A, Cikara M, Crockett MJ, Crum AJ, Douglas KM, Druckman JN, Drury J. Using social and behavioural science to support COVID-19 pandemic response. Nature Human Behaviour. 2020 Apr 30:1-2. 895

[49] M Dur-e-Ahmad, M Imran, Transmission Dynamics Model of Coronavirus COVID19 for the Outbreak in Most Affected Countries of the World, International Journal of Interactive Multimedia and Artificial Intelligence 6(2), pp. 7-10, 2020

[50] Wang Y, McKee M, Torbica A, Stuckler D. Systematic literature review on the spread of health-related misinformation on social media. Social Science & Medicine. 900 2019 Sep 18:112552.

[51] Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. Psychological science, 0956797620939054.

[52] Latif, Siddique; Usman, Muhammad; Manzoor, Sanaullah; Iqbal, Waleed; Qadir, 905 Junaid; Tyson, Gareth; et al. (2020): Leveraging Data Science To Combat COVID-Journal Pre-proof Journal Pre-proof19: A Comprehensive Review. TechRxiv. Preprint.

[53] Saiz, F., & Barandiaran, I. (2020). COVID-19 Detection in Chest X-ray Images using a Deep Learning Approach. International Journal Of Interactive Multimedia And 910 Artificial Intelligence, 6(Regular Issue), 4.

[54] Samuel, J.; Ali, G.G.M.N.; Rahman, M.M.; Esawi, E.; Samuel, Y. COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification. Information 2020, 11, 314.

[55] N2- Barkur, G., Vibha, & Kamath, G. B. (2020). Sentiment analysis of nationwide 915 lockdown due to COVID 19 outbreak: Evidence from India. Asian journal of psychiatry, 51, 102089. Advance online publication.

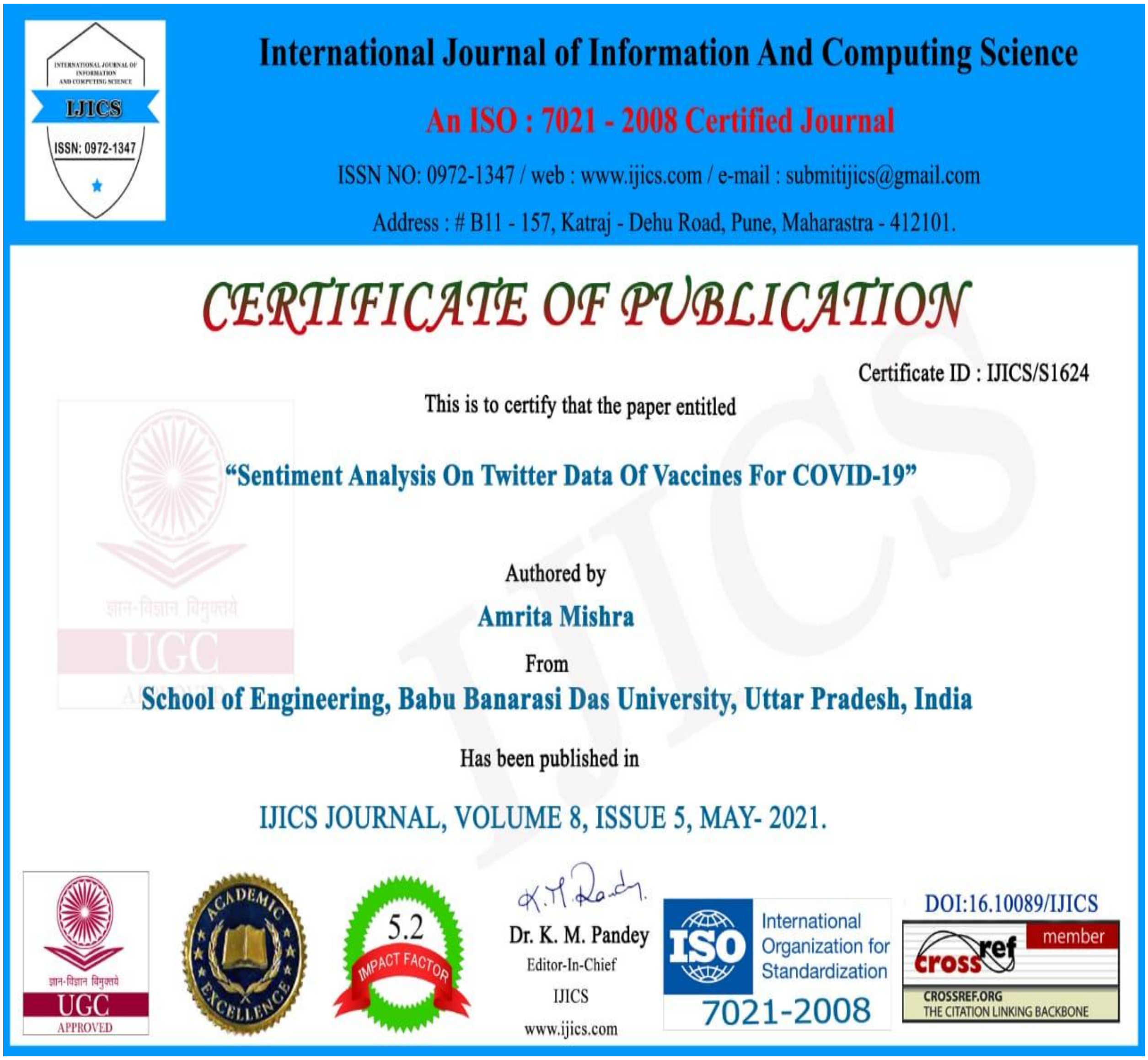
[56] Binti Hamzah FA, Lau C, Nazri H, Ligot DV, Lee G, Tan CL, et al. CoronaTracker: Worldwide COVID-19 Outbreak Data Analysis and Prediction. [Preprint]. Bull World 920 Health Organ. E-pub: 19 March 2020.

[57] Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study J Med Internet Res 2020;22(4):e19016DOI: 10.2196/19016 PMID: 32287039.

[58] Li, S., Wang, Y., Xue, J., Zhao, N., & Zhu, T. (2020). The impact of COVID-19 925 epidemic declaration on psychological consequences: a study on active Weibo users. International journal of environmental research and public health, 17(6), 2032.

[59] HuiDS IA, Madani TA, Ntoumi F, Koch R, Dar O. The continuing 2019-nCoV epidemic threat of novel corona viruses to global health: the latest 2019 novel coronavirus outbreak in Wuhan, China. Int J Infect Dis. 2020; 91:264-6

# Appendix 1`



INTERNATIONAL JOURNAL OF INFORMATION AND COMPUTING SCIENCE

IJICS

ISSN: 0972-1347

International Journal of Information And Computing Science

An ISO : 7021 - 2008 Certified Journal

ISSN NO: 0972-1347 / web : www.ijics.com / e-mail : submitijics@gmail.com

Address : # B11 - 157, Katraj - Dehu Road, Pune, Maharastra - 412101.

## CERTIFICATE OF PUBLICATION

Certificate ID : IJICS/S1624

This is to certify that the paper entitled

**“Sentiment Analysis On Twitter Data Of Vaccines For COVID-19”**

Authored by

**Amrita Mishra**

From

**School of Engineering, Babu Banarasi Das University, Uttar Pradesh, India**

Has been published in

IJICS JOURNAL, VOLUME 8, ISSUE 5, MAY- 2021.

ज्ञान-विज्ञान विमुक्तये
UGC
APPROVED

ACADEMIC EXCELLENCE

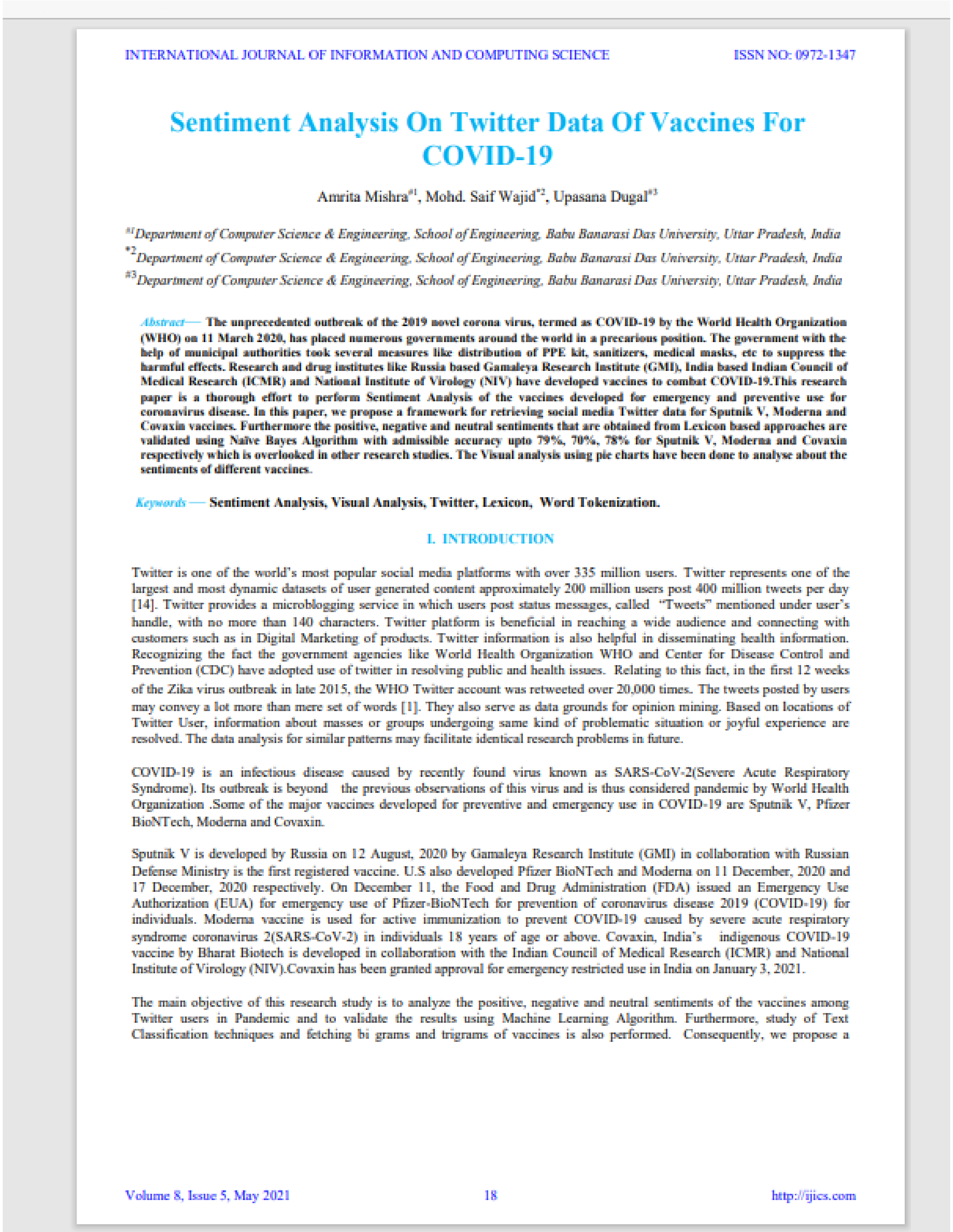
5.2
IMPACT FACTOR

Dr. K. M. Pandey
Editor-In-Chief
IJICS
www.ijics.com

ISO
International Organization for Standardization
7021-2008

DOI:16.10089/IJICS

crossref member
CROSSREF.ORG
THE CITATION LINKING BACKBONE



# Sentiment Analysis On Twitter Data Of Vaccines For COVID-19

Amrita Mishra[#1], Mohd. Saif Wajid[*2], Upasana Dugal[#3]

[#1]*Department of Computer Science & Engineering, School of Engineering, Babu Banarasi Das University, Uttar Pradesh, India*

[*2]*Department of Computer Science & Engineering, School of Engineering, Babu Banarasi Das University, Uttar Pradesh, India*

[#3]*Department of Computer Science & Engineering, School of Engineering, Babu Banarasi Das University, Uttar Pradesh, India*

*Abstract*— **The unprecedented outbreak of the 2019 novel corona virus, termed as COVID-19 by the World Health Organization (WHO) on 11 March 2020, has placed numerous governments around the world in a precarious position. The government with the help of municipal authorities took several measures like distribution of PPE kit, sanitizers, medical masks, etc to suppress the harmful effects. Research and drug institutes like Russia based Gamaleya Research Institute (GMI), India based Indian Council of Medical Research (ICMR) and National Institute of Virology (NIV) have developed vaccines to combat COVID-19.This research paper is a thorough effort to perform Sentiment Analysis of the vaccines developed for emergency and preventive use for coronavirus disease. In this paper, we propose a framework for retrieving social media Twitter data for Sputnik V, Moderna and Covaxin vaccines. Furthermore the positive, negative and neutral sentiments that are obtained from Lexicon based approaches are validated using Naïve Bayes Algorithm with admissible accuracy upto 79%, 70%, 78% for Sputnik V, Moderna and Covaxin respectively which is overlooked in other research studies. The Visual analysis using pie charts have been done to analyse about the sentiments of different vaccines.**

*Keywords* — **Sentiment Analysis, Visual Analysis, Twitter, Lexicon, Word Tokenization.**

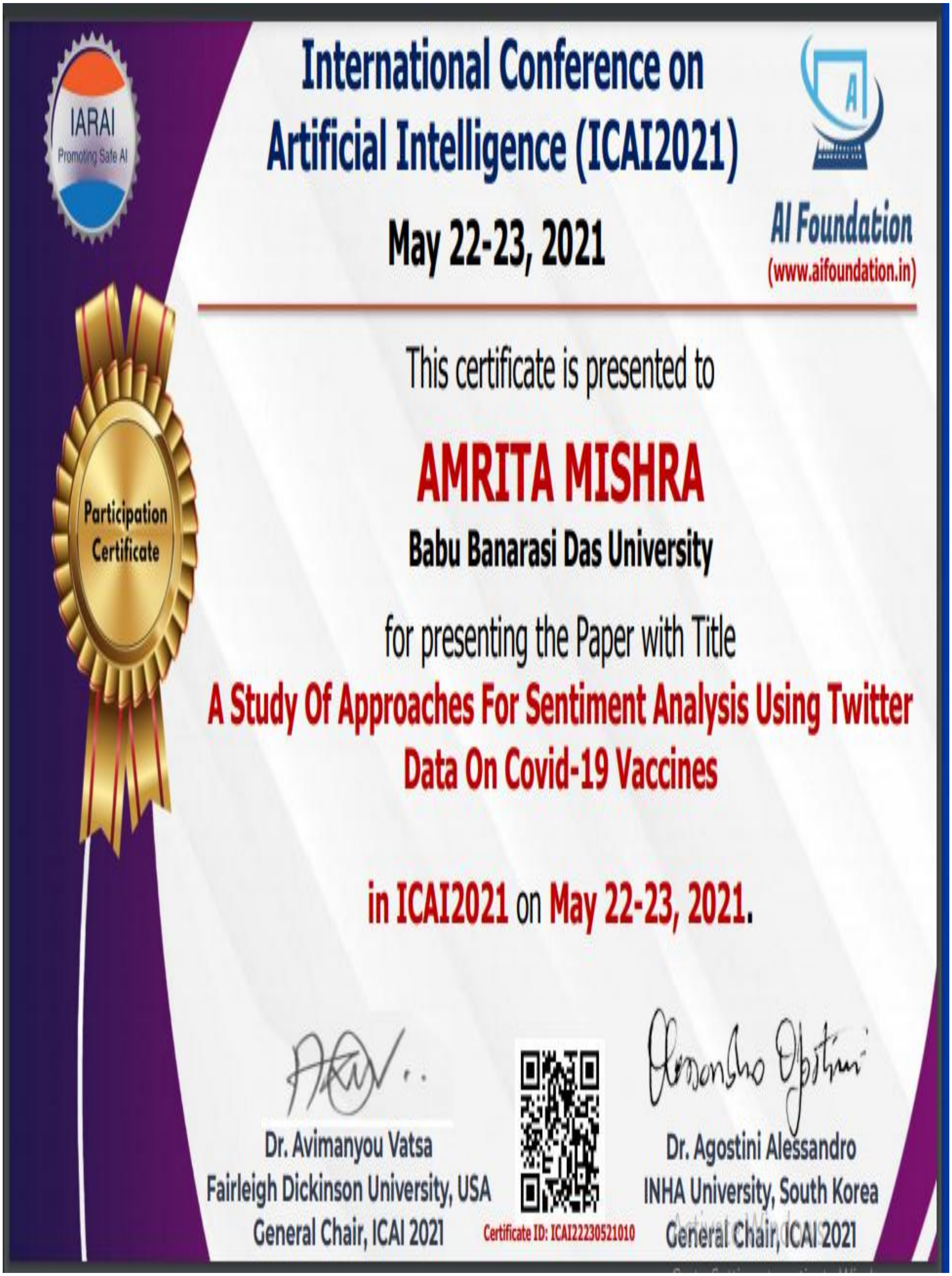
## I. INTRODUCTION

Twitter is one of the world's most popular social media platforms with over 335 million users. Twitter represents one of the largest and most dynamic datasets of user generated content approximately 200 million users post 400 million tweets per day [14]. Twitter provides a microblogging service in which users post status messages, called "Tweets" mentioned under user's handle, with no more than 140 characters. Twitter platform is beneficial in reaching a wide audience and connecting with customers such as in Digital Marketing of products. Twitter information is also helpful in disseminating health information. Recognizing the fact the government agencies like World Health Organization WHO and Center for Disease Control and Prevention (CDC) have adopted use of twitter in resolving public and health issues. Relating to this fact, in the first 12 weeks of the Zika virus outbreak in late 2015, the WHO Twitter account was retweeted over 20,000 times. The tweets posted by users may convey a lot more than mere set of words [1]. They also serve as data grounds for opinion mining. Based on locations of Twitter User, information about masses or groups undergoing same kind of problematic situation or joyful experience are resolved. The data analysis for similar patterns may facilitate identical research problems in future.

COVID-19 is an infectious disease caused by recently found virus known as SARS-CoV-2(Severe Acute Respiratory Syndrome). Its outbreak is beyond the previous observations of this virus and is thus considered pandemic by World Health Organization .Some of the major vaccines developed for preventive and emergency use in COVID-19 are Sputnik V, Pfizer BioNTech, Moderna and Covaxin.

Sputnik V is developed by Russia on 12 August, 2020 by Gamaleya Research Institute (GMI) in collaboration with Russian Defense Ministry is the first registered vaccine. U.S also developed Pfizer BioNTech and Moderna on 11 December, 2020 and 17 December, 2020 respectively. On December 11, the Food and Drug Administration (FDA) issued an Emergency Use Authorization (EUA) for emergency use of Pfizer-BioNTech for prevention of coronavirus disease 2019 (COVID-19) for individuals. Moderna vaccine is used for active immunization to prevent COVID-19 caused by severe acute respiratory syndrome coronavirus 2(SARS-CoV-2) in individuals 18 years of age or above. Covaxin, India's indigenous COVID-19 vaccine by Bharat Biotech is developed in collaboration with the Indian Council of Medical Research (ICMR) and National Institute of Virology (NIV).Covaxin has been granted approval for emergency restricted use in India on January 3, 2021.

The main objective of this research study is to analyze the positive, negative and neutral sentiments of the vaccines among Twitter users in Pandemic and to validate the results using Machine Learning Algorithm. Furthermore, study of Text Classification techniques and fetching bi grams and trigrams of vaccines is also performed. Consequently, we propose a



IARAI
Promoting Safe AI

# International Conference on Artificial Intelligence (ICAI2021)

May 22-23, 2021

AI Foundation
(www.aifoundation.in)

Participation Certificate

This certificate is presented to

**AMRITA MISHRA**

**Babu Banarasi Das University**

for presenting the Paper with Title

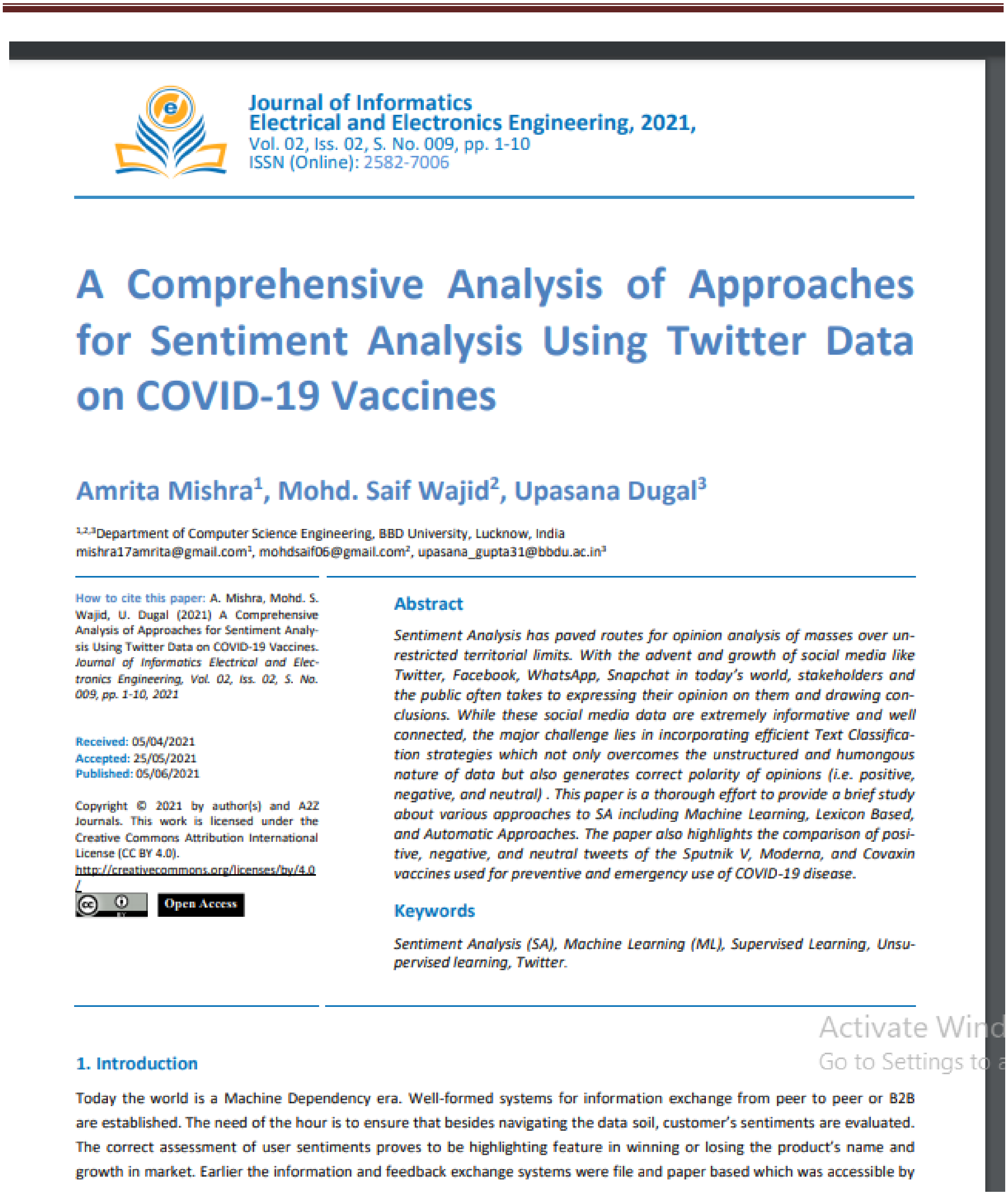
**A Study Of Approaches For Sentiment Analysis Using Twitter Data On Covid-19 Vaccines**

**in ICAI2021** on **May 22-23, 2021.**

Dr. Avimanyou Vatsa
Fairleigh Dickinson University, USA
General Chair, ICAI 2021

Certificate ID: ICAI22230521010

Dr. Agostini Alessandro
INHA University, South Korea
General Chair, ICAI 2021



Journal of Informatics
Electrical and Electronics Engineering, 2021,
Vol. 02, Iss. 02, S. No. 009, pp. 1-10
ISSN (Online): 2582-7006

# A Comprehensive Analysis of Approaches for Sentiment Analysis Using Twitter Data on COVID-19 Vaccines

Amrita Mishra[1], Mohd. Saif Wajid[2], Upasana Dugal[3]

[1,2,3]Department of Computer Science Engineering, BBD University, Lucknow, India
mishra17amrita@gmail.com[1], mohdsaif06@gmail.com[2], upasana_gupta31@bbdu.ac.in[3]

How to cite this paper: A. Mishra, Mohd. S. Wajid, U. Dugal (2021) A Comprehensive Analysis of Approaches for Sentiment Analysis Using Twitter Data on COVID-19 Vaccines. *Journal of Informatics Electrical and Electronics Engineering, Vol. 02, Iss. 02, S. No. 009, pp. 1-10, 2021*

Received: 05/04/2021
Accepted: 25/05/2021
Published: 05/06/2021

## Abstract

*Sentiment Analysis has paved routes for opinion analysis of masses over unrestricted territorial limits. With the advent and growth of social media like Twitter, Facebook, WhatsApp, Snapchat in today's world, stakeholders and the public often takes to expressing their opinion on them and drawing conclusions. While these social media data are extremely informative and well connected, the major challenge lies in incorporating efficient Text Classification strategies which not only overcomes the unstructured and humongous nature of data but also generates correct polarity of opinions (i.e. positive, negative, and neutral) . This paper is a thorough effort to provide a brief study about various approaches to SA including Machine Learning, Lexicon Based, and Automatic Approaches. The paper also highlights the comparison of positive, negative, and neutral tweets of the Sputnik V, Moderna, and Covaxin vaccines used for preventive and emergency use of COVID-19 disease.*

## Keywords

*Sentiment Analysis (SA), Machine Learning (ML), Supervised Learning, Unsupervised learning, Twitter.*

## 1. Introduction

Today the world is a Machine Dependency era. Well-formed systems for information exchange from peer to peer or B2B are established. The need of the hour is to ensure that besides navigating the data soil, customer's sentiments are evaluated. The correct assessment of user sentiments proves to be highlighting feature in winning or losing the product's name and growth in market. Earlier the information and feedback exchange systems were file and paper based which was accessible by

# Appendix 2

## Sample Code

```
# Import Libraries
from textblob import TextBlob
import sys
import tweepy
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import os
import nltk
import pycountry
import re
import string
from wordcloud import WordCloud, STOPWORDS
from PIL import Image
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from langdetect import detect
from nltk.stem import SnowballStemmer
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from sklearn.feature_extraction.text import CountVectorizer
#Authorization and Search tweets
#Getting authorization
consumer_key = 'AdEa2Uo4mFnzOJQwbicfOpcyI'
consumer_key_secret = 'r4NvJEQFEVB1ufFj3W7IBQls4O2gPJM27Dz4248Y4X05im8xNl'
access_token = '2579074550-FFVrSTR2a7B51CLlNbuq6m3A3g37X22LLdiQxhS'
access_token_secret = '********************************************6'
auth = tweepy.OAuthHandler(consumer_key, consumer_key_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth, wait_on_rate_limit=True)
```

```
print("successful")

def percentage(part,whole):
    return 100 * float(part)/float(whole)
positive = 0
negative = 0
neutral = 0
polarity = 0
tweet_list = []
neutral_list = []
negative_list = []
positive_list = []

for tweet in tweets:

    #print(tweet.text)
    tweet_list.append(tweet.text)
    analysis = TextBlob(tweet.text)
    score = SentimentIntensityAnalyzer().polarity_scores(tweet.text)
    neg = score['neg']
    neu = score['neu']
    pos = score['pos']
    comp = score['compound']
    polarity += analysis.sentiment.polarity

    if neg > pos:
        negative_list.append(tweet.text)
        negative += 1

    elif pos > neg:
        positive_list.append(tweet.text)
        positive += 1

    elif pos == neg:
        neutral_list.append(tweet.text)
        neutral += 1

positive = percentage(positive, max_tweets)
negative = percentage(negative, max_tweets)
neutral = percentage(neutral, max_tweets)
polarity = percentage(polarity, max_tweets)
positive = float(format(positive, '.1f'))
negative = float(format(negative, '.1f'))
neutral = float(format(neutral, '.1f'))
print('successful')
tweet_list
 #Appliyng tokenization
def tokenization(text):
    text = re.split('\W+', text)
    return text
tw_list['tokenized'] = tw_list['punct'].apply(lambda x: tokenization(x.lower()))
stopword = nltk.corpus.stopwords.words('english')
def remove_stopwords(text):
    text = [word for word in text if word not in stopword]
    return text

tw_list['nonstop'] = tw_list['tokenized'].apply(lambda x: remove_stopwords(x))
#Appliyng Stemmer
ps = nltk.PorterStemmer()

#Appliyng Countvectorizer
countVectorizer = CountVectorizer(analyzer=clean_text)
```

```
countVector = countVectorizer.fit_transform(tw_list['text'])
print('{} Number of reviews has {} words'.format(countVector.shape[0], countVector.shape[1]))
#print(countVectorizer.get_feature_names())
# Most Used Words
count = pd.DataFrame(count_vect_df.sum())
countdf = count.sort_values(0,ascending=False).head(20)
countdf[1:11] #Function to ngram
def get_top_n_gram(corpus,ngram_range,n=None):
    vec = CountVectorizer(ngram_range=ngram_range,stop_words = 'english').fit(corpus)
    bag_of_words = vec.transform(corpus)
    sum_words = bag_of_words.sum(axis=0)
    words_freq = [(word, sum_words[0, idx]) for word, idx in vec.vocabulary_.items()]
    words_freq =sorted(words_freq, key = lambda x: x[1], reverse=True)
    return words_freq[:n]
#n2_bigram
n2_bigrams = get_top_n_gram(tw_list['text'],(2,2),20)

n2_bigrams
#n3_trigram
n3_trigrams = get_top_n_gram(tw_list['text'],(3,3),20)

n3_trigrams
#split the dataset
#train dataset
train_text=df.text[:300]
train_sentiment=df.sentiment[:300]
#test dataset
test_text=df.text[802:]
test_sentiments=df.sentiment[802:]
print(train_text.shape,train_sentiment.shape)
print(test_text.shape,test_sentiments.shape)
```

```
X_train = df.loc[:700, 'text'].values
y_train = df.loc[:700, 'sentiment'].values
X_test = df.loc[300:, 'text'].values
y_test = df.loc[300:, 'sentiment'].values
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer()
train_vectors = vectorizer.fit_transform(X_train)
test_vectors = vectorizer.transform(X_test)
print(train_vectors.shape, test_vectors.shape)
from sklearn.naive_bayes import MultinomialNB
clf = MultinomialNB().fit(train_vectors, y_train)
from  sklearn.metrics  import accuracy_score
predicted = clf.predict(test_vectors)
print(accuracy_score(y_test,predicted))
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score
from sklearn.metrics import precision_recall_curve
from sklearn.metrics import auc
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
from sklearn.metrics import classification_report
print(classification_report(y_test, predicted))
```

**BBDU-PG-FORM 01**

## BABU BANARASI DAS UNIVERSITY, LUCKNOW

## CERTIFICATE OF FINALTHESIS SUBMISSION

(To be submitted in duplicate)

1.Name: ……AMRITAMISHRA…………………………………………………………......................

2.EnrollmentNo.…… 11904490637…………………………………………………..………………

3. Thesis title: …SENTIMENT ANALYSIS ON TWITTER DATA OF VACCINES FOR COVID-19………………………………………..…………………………………………

4. Degree for which the thesis is submitted: ……MASTER OF TECHNOLOGY………..................

5. School (of the University to which the thesis is submitted)

….BABU BANARASI DAS UNIVERSITY ,LUCKNOW…………………………………….

6. Thesis Preparation Guide was referred to for preparing the thesis. ✓YES NO

7. Specifications regarding thesis format have been closely followed. ✓YES NO

8. The contents of the thesis have been organized based on the guidelines. ✓YES NO

9. The thesis has been prepared without resorting to plagiarism. ✓YES NO

10. All sources used have been cited appropriately. ✓ YES NO

11. The thesis has not been submitted elsewhere for a degree. ✓YES NO

12. All the corrections have been incorporated. ✓YES NO

13. Submitted 4 hard bound copies plus one CD. ✓YES NO

(Signature(s) of the Supervisor(s)) (Signature of the Candidate)

Name(s): Mr. Mohd. Saif Wajid
Mrs. Upasana Dugal

NAME: Amrita Mishra
ROLL NO: 1190449001
ENROLLMENT NO.:11904490637

upasana

**BBDU-PG-FORM 02**

**BABU BANARASI DAS UNIVERSITY, LUCKNOW**

**CERTIFICATE OF THESIS SUBMISSION FOR EVALUATION**

(Submit in Duplicate)

1. Name: ………AMRITA MISHRA…………………………………………………………..

2. Enrollment No.:…11904490637……………………………………………………………

3.Thesis title:..SENTIMENT ANALYSIS ON TWITTERDATAOFVACCINES FOR…COVID19……………..………..……………………………………...……………

4. Degree for which the thesis is submitted: … MASTER OF TECHNOLOGY………………

5. Faculty of the University to which the thesis is submitted
….PROF. MOHD. SAIF WAJID AND PROF.UPASANA DUGAL……………...………..

6. Thesis Preparation Guide was referred to for preparing the thesis. ✓ YES NO

7. Specifications regarding thesis format have been closely followed. ✓ YES NO

8. The contents of the thesis have been organized based on the guidelines. ✓ YES NO

9. The thesis has been prepared without resorting to plagiarism. ✓ YES NO

10. All sources used have been cited appropriately. ✓ YES NO

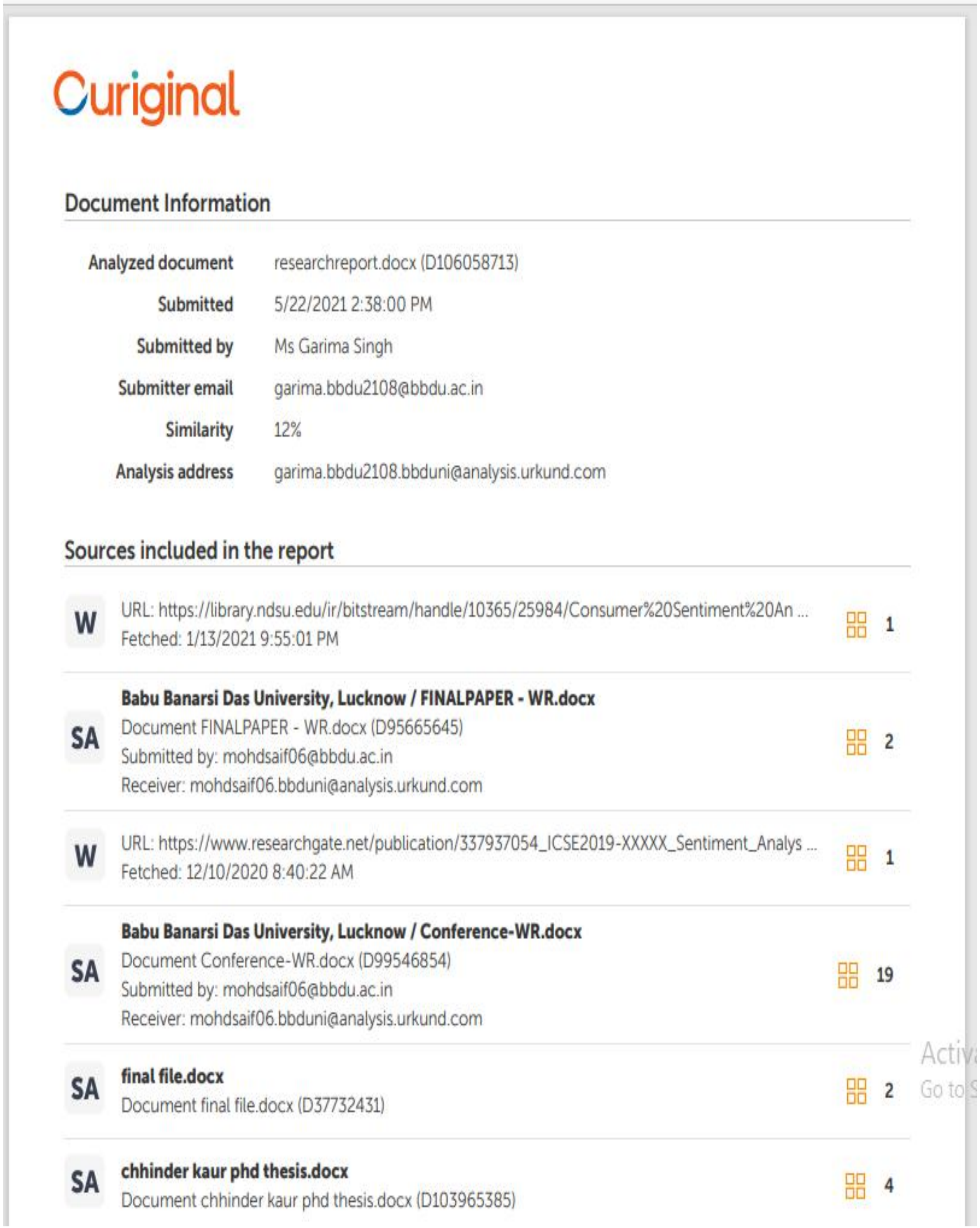
11. The thesis has not been submitted elsewhere for a degree. ✓ YES NO

12. Submitted 2 spiral bound copies plus one CD. ✓ YES NO

(Signature of the Candidate)

MISHRA AMRITA

Name…AMRITA MISHRA……………
Roll No …1190449001………..…………
Enrollment No.:… 11904490637………..



Curiginal

## Document Information

| | |
|---|---|
| **Analyzed document** | researchreport.docx (D106058713) |
| **Submitted** | 5/22/2021 2:38:00 PM |
| **Submitted by** | Ms Garima Singh |
| **Submitter email** | garima.bbdu2108@bbdu.ac.in |
| **Similarity** | 12% |
| **Analysis address** | garima.bbdu2108.bbduni@analysis.urkund.com |

## Sources included in the report

| | | |
|---|---|---|
| W | URL: https://library.ndsu.edu/ir/bitstream/handle/10365/25984/Consumer%20Sentiment%20An ...<br>Fetched: 1/13/2021 9:55:01 PM | 1 |
| SA | **Babu Banarsi Das University, Lucknow / FINALPAPER - WR.docx**<br>Document FINALPAPER - WR.docx (D95665645)<br>Submitted by: mohdsaif06@bbdu.ac.in<br>Receiver: mohdsaif06.bbduni@analysis.urkund.com | 2 |
| W | URL: https://www.researchgate.net/publication/337937054_ICSE2019-XXXXX_Sentiment_Analys ...<br>Fetched: 12/10/2020 8:40:22 AM | 1 |
| SA | **Babu Banarsi Das University, Lucknow / Conference-WR.docx**<br>Document Conference-WR.docx (D99546854)<br>Submitted by: mohdsaif06@bbdu.ac.in<br>Receiver: mohdsaif06.bbduni@analysis.urkund.com | 19 |
| SA | **final file.docx**<br>Document final file.docx (D37732431) | 2 |
| SA | **chhinder kaur phd thesis.docx**<br>Document chhinder kaur phd thesis.docx (D103965385) | 4 |