SENTIMENTAL ANALYSIS ON COVID VACCINE TWEET DATA

A Project Report submitted in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING

Submitted by

Venkata Harshith Kamisetty,121810313024

Harshit Yellanti, 121810313031

Saiteja Potluru, 121810313035

PVV Surya Prasad, 121810313036

Under the esteemed guidance of

Mr. Durga Prasad B

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

GITAM (Deemed to be University)

VISAKHAPATNAM

November, 2021

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GITAM INSTITUTE OF TECHNOLOGY

GITAM

(Deemed to be University)

DECLARATION

I/We, hereby declare that the project report entitled "SENTIMENTAL ANALYSIS ON COVID VACCINE TWEET DATA" is an original work done in the Department of Computer Science and Engineering, GITAM Institute of Technology, GITAM (Deemed to be University) submitted in partial fulfillment of the requirements for the award of the degree of B.Tech. in Computer Science and Engineering.

The work has not been submitted to any other college or University for the award of any degree or diploma.

Date: 8 November, 2021

V. Harshith Kamisetty	121810313024
Harshit Yellanti	121810313031
Saiteja Potluru	121810313035
PVV Surya Prasad	121810313036

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GITAM INSTITUTE OF TECHNOLOGY GITAM

(Deemed to be University)

CERTIFICATE

This is to certify that the project report entitled "SENTIMENTAL ANALYSIS ON COVID VACCINE TWEET DATA" is a bonafide record of work carried out by Venkata Harshith Kamisetty, (121810313024), Harshit Yellanti (121810313031), Saiteja Potluru (121810313035), PVV Surya Prasad (121810313036) students submitted in partial fulfillment of requirement for the award of degree of Bachelors ofTechnology in Computer Science and Engineering.

Project Guide

Mr. Durga Prasad B
Assistant Professor

Head of the Department

Dr. R.Sireesha
Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GITAM INSTITUTE OF TECHNOLOGY GITAM

(Deemed to be University)

CERTIFICATE

This is to certify that the project report entitled "SENTIMENTAL ANALYSIS ON COVID VACCINE TWEET DATA" is a bonafide record of work carried out by Venkata Harshith Kamisetty, (121810313024), Harshit Yellanti (121810313031), Saiteja Potluru (121810313035), PVV Surya Prasad (121810313036) students submitted in partial fulfillment of requirement for the award of degree of Bachelors ofTechnology in Computer Science and Engineering.

Project Guide **Head of the Department**

Mr. Durga Prasad B **Dr. R.Sireesha**

Assistant Professor **Professor**

TABLE OF CONTENTS

Contents

SENTIMENTAL ANALYSIS ON COVID VACCINE TWEET DATA

1. ABSTRACT

The SARS-CoV-2 coronavirus disease (COVID-19) pandemic has a devastating effect on people and economics with many causalities and continues to impact the health and well-being of the global population. Pandemic has shifted our view of the world to a different dimension. The efficient way to control the increasing spread of COVID-19 lies in vaccinating the public. The Government has initiated the vaccination drives. However, there is uncertainty in the minds of the people regarding vaccines.

Ever since the vaccination drive for COVID-19 started in India, the citizens have been sharing their views on social media every day. These views are generally required to analyze and detect the general opinion and enhance the decision-making process. In this project, we will analyze the public tweets from Twitter related to COVID-19 vaccinations to detect the user's opinion on the particular vaccine. We use Robotic Process Automation(RPA) to collect the Twitter data and Machine Learning (ML) Algorithms like TextBlob, NLTK Vader to project the sentimental analysis.

2. INTRODUCTION

A pandemic is considered a phenomenon having devastating effects on people and economics with many causalities. Pandemics are said to have both health and economic calamities. The spread of the deadly virus highlights the importance of vaccination at a national level. The vaccination is supposed to protect the nation from continued damage.

The current response to the COVID-19 pandemic involves the aggressive implementation of containment, suppression, and mitigation strategies causing devastating social, economic, and political crises. The restrictions on our lives are the only thing holding the virus in check. The world cannot return to normal without safe and effective vaccines against COVID-19. As the vaccination drives have started, many people are confused about choosing the best available vaccines. As we know the importance of social media in making decisions, People express their views and thoughts daily. These views, which generally take the unstructured format, are analyzed to detect general opinions and enhance decision-making. As the COVID-19 vaccination process was initiated, people started expressing their views on the same. To analyze the same, we used the process of sentiment analysis along with Natural Language Processing. Sentimental analysis is a method of categorizing the sample texts into positive, negative, or neutral brackets. The sentiment score is obtained from the sentimental analysis. Every word in the statement, whether positive, negative, or neutral, contains a sentiment score. Using the score, the model will determine whether the particular tweet has positive, negative, or neutral sentiment.

3.LITERATURE REVIEW

There have been several works related to analysing the Twitter dataset on different topics during the COVID-19 pandemic .Only a few studies focus on the Twitter data related to COVID-19 vaccination .

In this literature review, it shows the process for collection and pre processing of data for better result. Praveen Sv, Jyoti Tandon, Vikas and Hitesh Hinduja(2021) [1]has proposed a process to extract the tweets posts selecting the tweets, the process of data cleaning takes place which aims at removing the punctuation, emoticons, images, hyperlinks, numbers, and stop words. Only "Text" shall be considered for analysis. Stop words needs to be filtered out as they have no meaning of their own and removing them from the sentence leaves the meaning of the sentence unaltered. They are not required for analysis.

Performing Sentimental analysis is a tricky task as there are many models to perform, the best models are TextBlob [6] Loria, is a Python library that provides support for different Natural Language Processing (NLP) tasks including sentiment analysis. TextBlob outputs the following two metrics for any input text.

VADER [7] Hutto, is a lexicon- and rule-based sentiment analysis tool. It is specifically designed for sentiments expressed on social media and works well on texts from other domains as well.

4.PROBLEM IDENTIFICATION

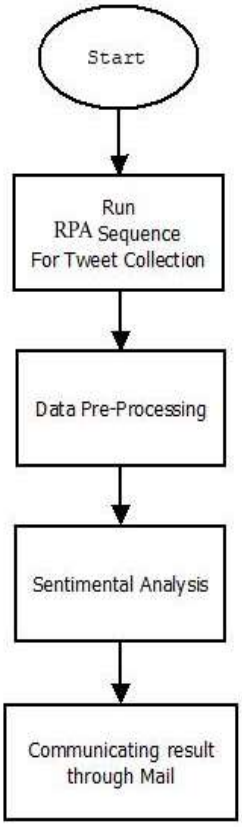
COVID-19 pandemic, is also known as the coronavirus pandemic, is an ongoing serious global problem all over the world. The outbreak first came to light in December 2019 in Wuhan, China. This was declared pandemic by the World Health Organization on 11th March 2020. Many governments took actions to prevent the spread of the virus. Some governments have misled their citizens by tuning down the severity of the virus. Covid vaccines are effective preventive measures against the virus. But some media has only highlighted the vaccine's side effects and negative views on them. People took social media platforms to share their emotions, and opinions during this lockdown to find a way to relax and calm down. This project analyses all the tweets tweeted by the people who took the vaccine.

4.1. OBJECTIVE

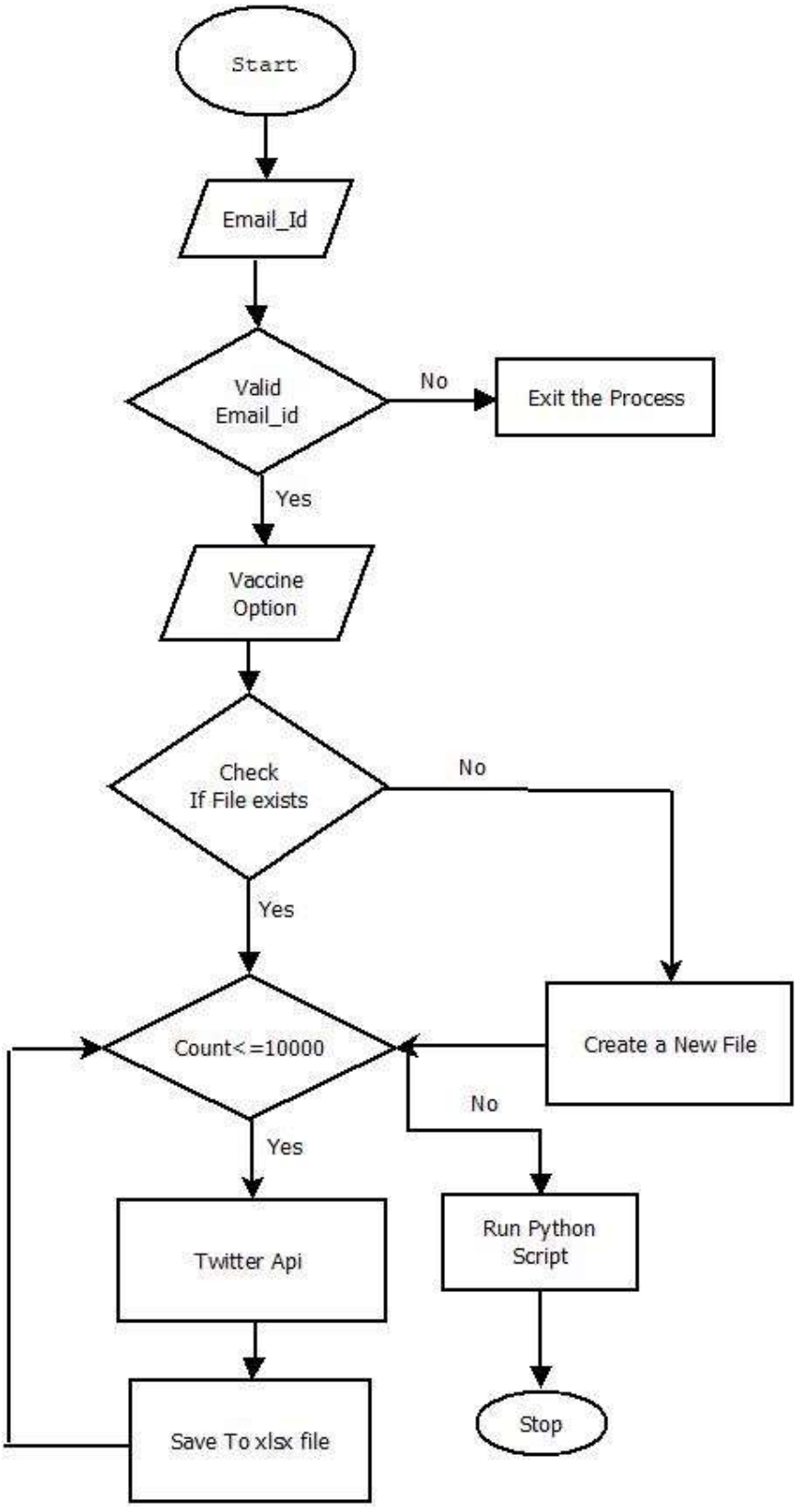
We identified public sentiments and opinions towards the COVID-19 vaccines based on the content of twitter data. The project mainly aims to provide analysis on various vaccines using tweets which are posted by various users, for which we are using Natural language processing (NLP) in particularly we are using lexicon based approach like Text Blob and Vader where each model gives a polarity score range from [-1,1] by this we can tell that whether a particular tweet is positive, neutral or negative and provide an analysis

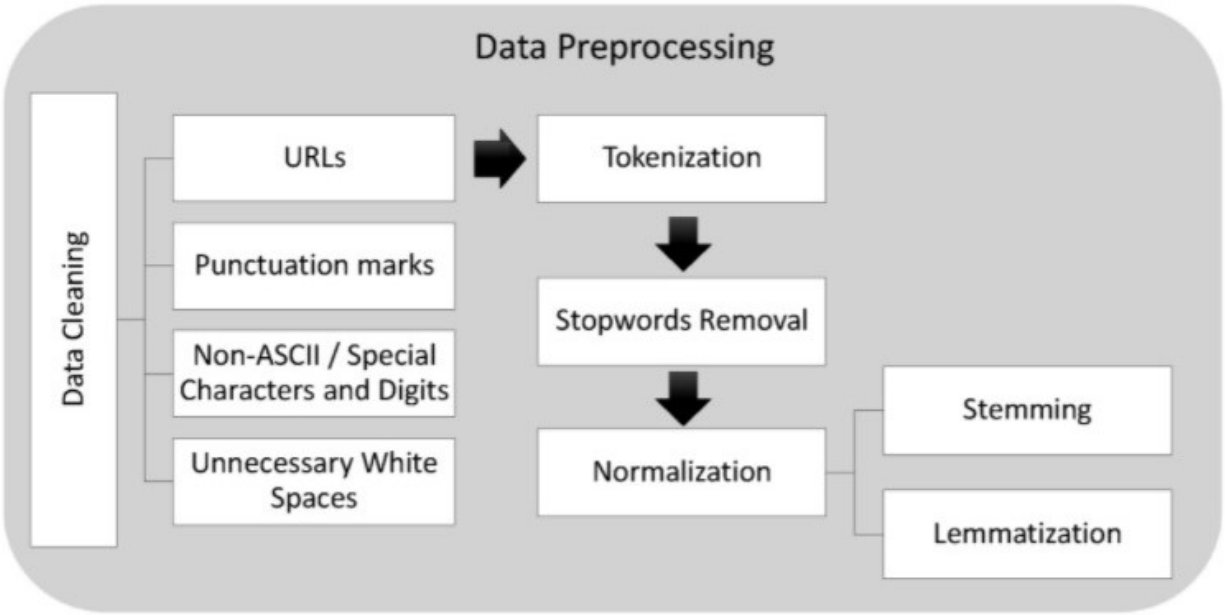
5. SYSTEM METHODOLOGY

The below flowchart depicts Flow of the Project



RPA FLOW CHART





5.1. WORKING OF THE PROJECT :

1) Run the sequence which was in the uipath

2) After validation of inputs it start executing the twitter api

3) Where with help of twitter search api we will get the desired number of tweets and then we will store them in a excel file.

4) After Successful collection of data we will perform a pre-process of the data.

5) Then, We perform sentimental analysis using 2 libraries know as text blob, Vader.

6) The final output of review about the vaccine will be sent to their respective mail ids.

6.OVERVIEW OF TECHNOLOGIES

1. VADER: In this project we have used a sentimental analyzer tool VADER (Valence Aware Dictionary and sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains. Vader only performs sentiment analysis on English texts, but that workaround (automatic translation) may be a viable option. As mainly it relies on a dictionary that maps lexical features to emotion intensities known as sentiment scores. The sentiment score of a text can be obtained by summing up the intensity of each word in the text.

It is available in the NLTK package and can be applied directly to unlabeled text data**.**

VADER's SentimentIntensityAnalyzer() takes in a string and returns a dictionary of scores in each of four categories:

- negative
- neutral
- positive
- compound *(computed by normalizing the scores above*

The Compound score is a metric that calculates the sum of all the lexicon ratings which have been normalized between -1(most extreme negative) and +1 (most extreme positive).

2.TEXTBLOB: TextBlob is a python library for Natural Language Processing (NLP).TextBlob actively used Natural Language ToolKit (NLTK) to achieve its tasks. TextBlob is a simple library which supports complex analysis and operations on textual data.

TextBlob returns polarity and subjectivity of a sentence. Polarity lies between [-1,1], -1 defines a negative sentiment and 1 defines a positive sentiment. Negation words reverse the polarity. Subjectivity lies between [0,1]. Subjectivity quantifies the amount of personal opinion and factual information contained in the text. The higher subjectivity means that the text contains personal opinion rather than factual information. TextBlob has one more parameter — intensity. TextBlob calculates subjectivity by looking at the 'intensity'. Intensity determines if a word modifies the next word. For English, adverbs are used as modifiers ('very good').

3.RPA: Robotic process automation (RPA) is a software technology that makes it easy to build, deploy, and manage software robots that emulate humans actions interacting with digital systems and software

Just like people, software robots can do things like understand what's on a screen, complete the right keystrokes, navigate systems, identify and extract data, and perform a wide range of defined actions. But software robots can do it faster and more consistently than people, without the need to get up and stretch or take a coffee break.

In this project we are using RPA for collection of tweets from twitter api.

4.LEMMATIZATION: Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item. Lemmatization is similar to stemming but it brings context to the words. So it links words with similar meanings to one word.

Text preprocessing includes both Stemming as well as Lemmatization. Many times people find these two terms confusing. Some treat these two as the same. Actually, lemmatization is preferred over Stemming because lemmatization does morphological analysis of the words.

Applications of lemmatization are:

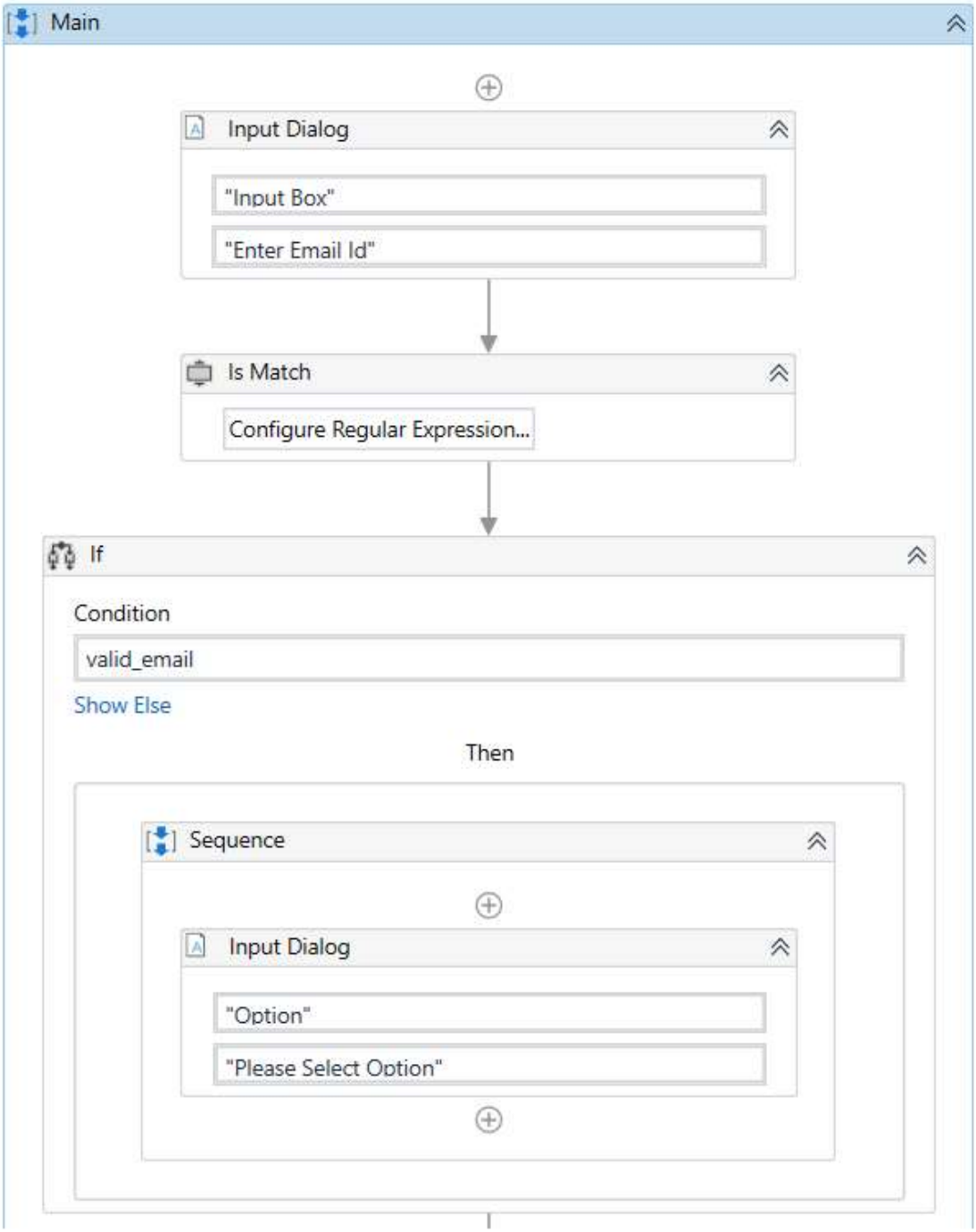
Used in comprehensive retrieval systems like search engines.

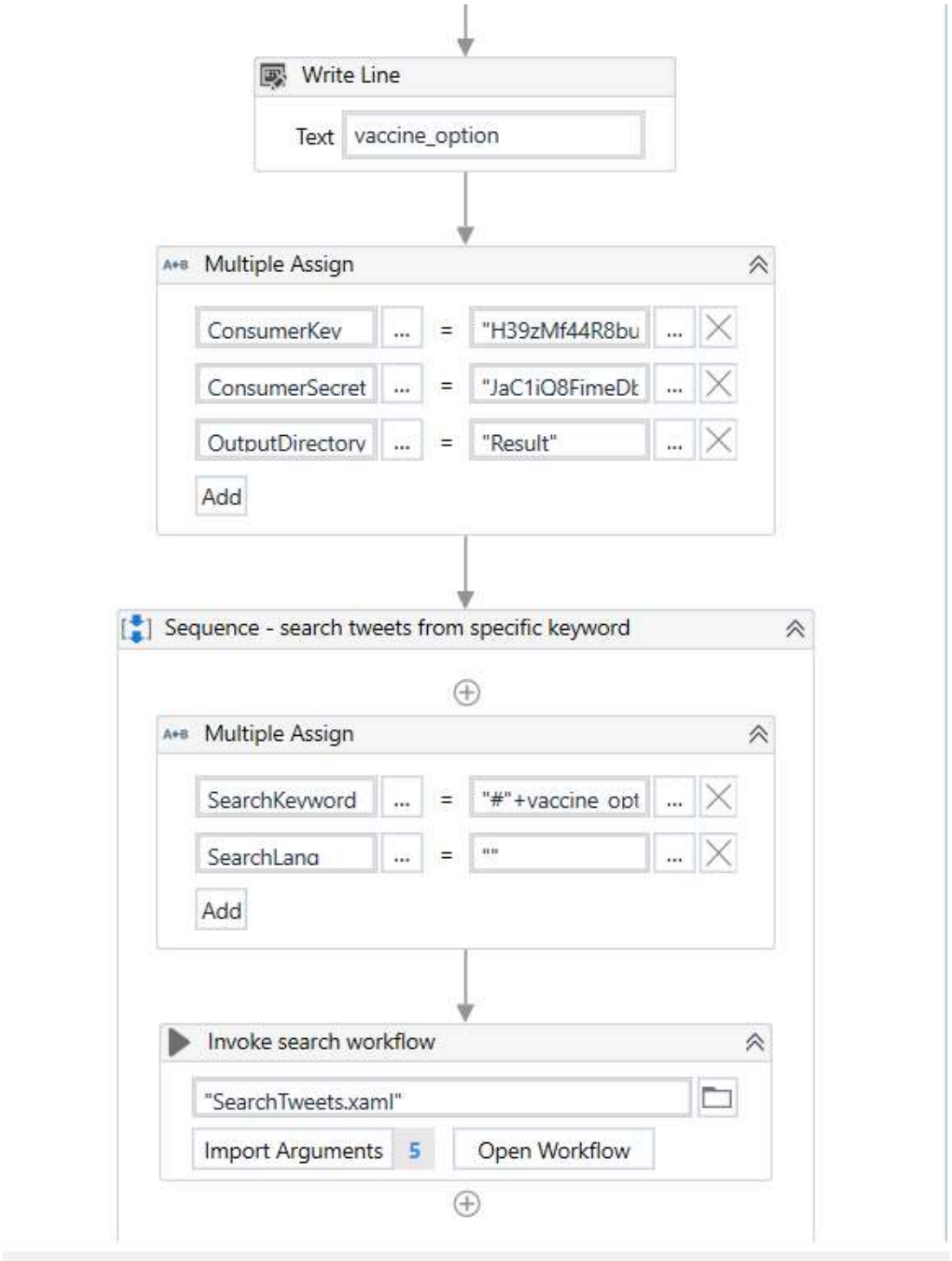
Used in compact indexing

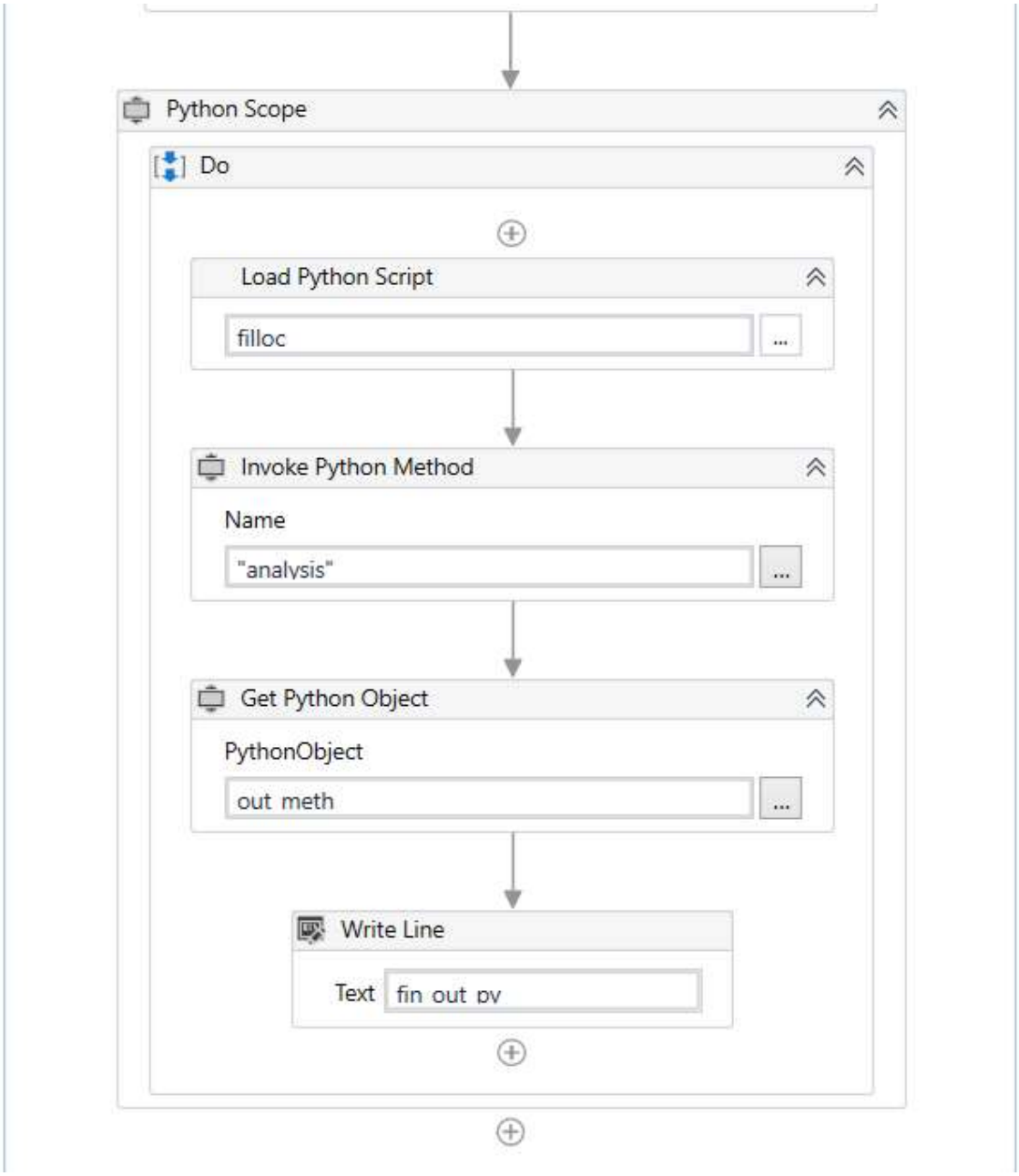
7. IMPLEMENTATION

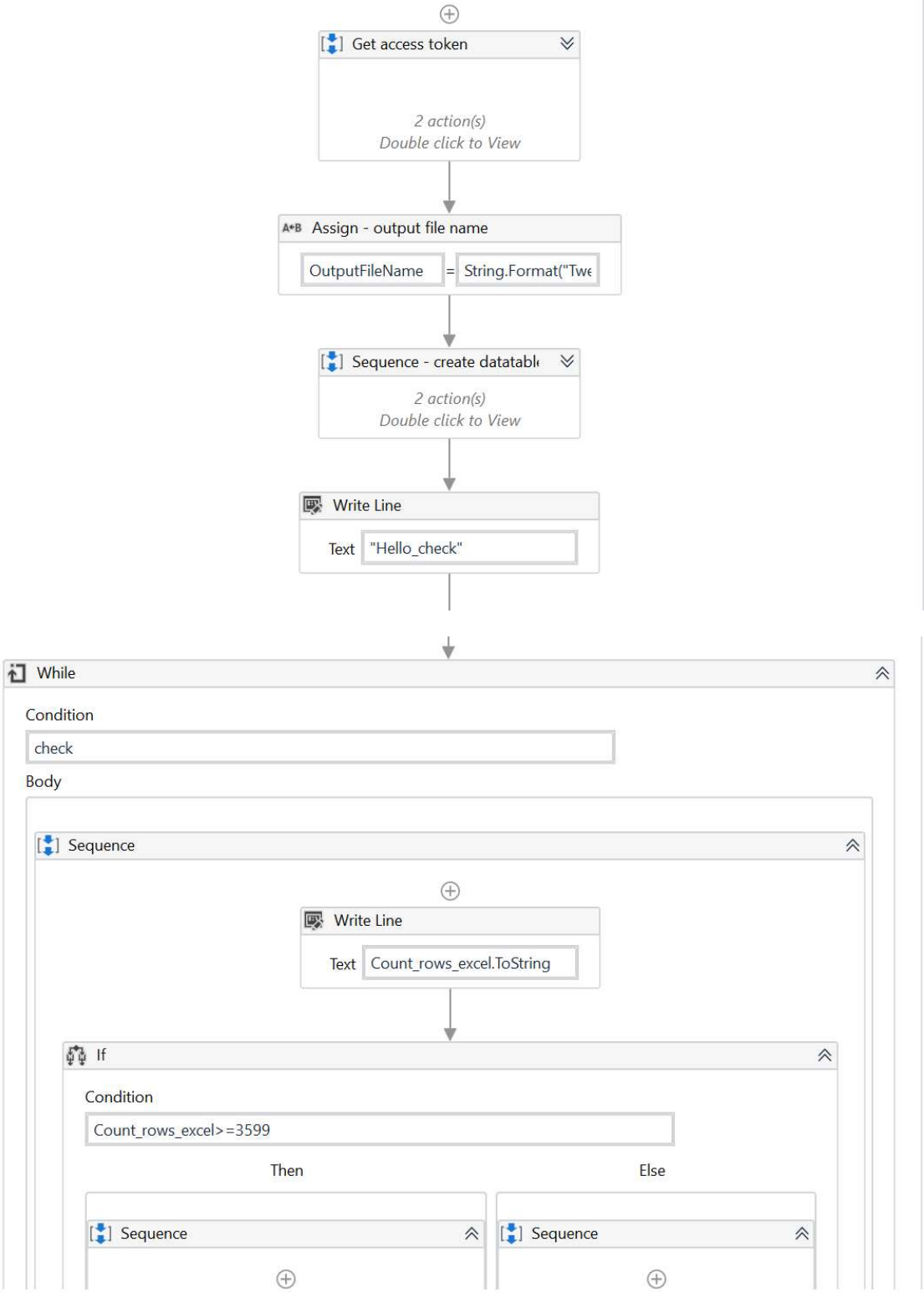
7.1 CODING:

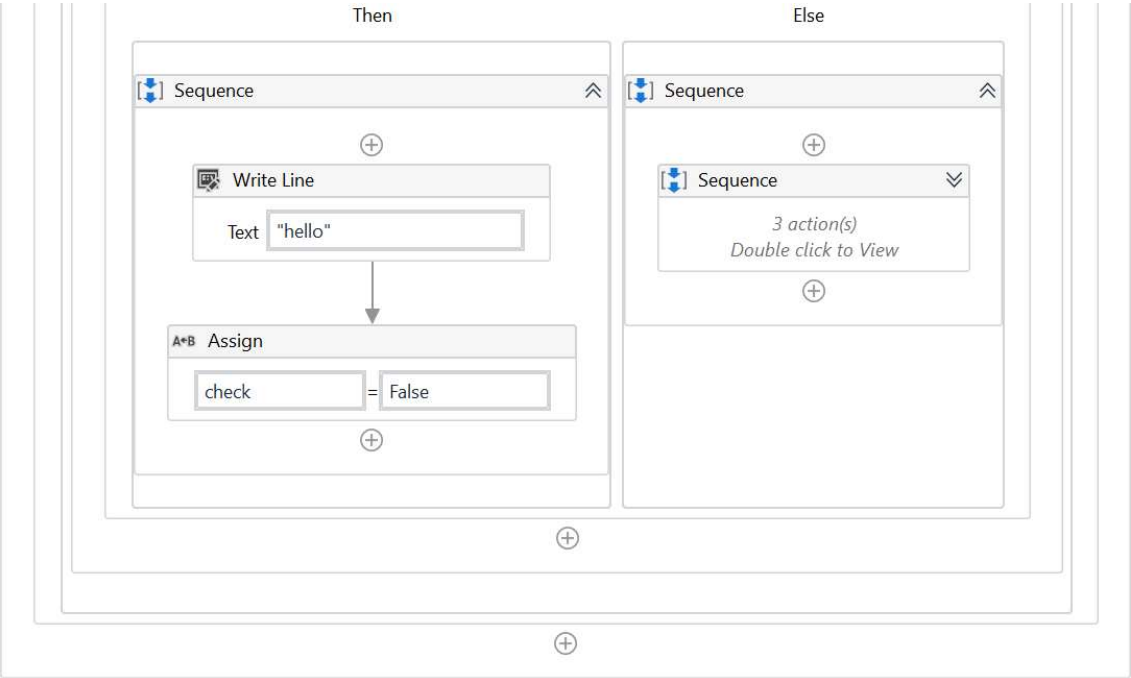
Sequence Diagram of Rpa




Write Line
Text
vaccine_option
Multiple Assign
ConsumerKey
=
"H39zMf44R8bu
ConsumerSecret
=
"JaC1iO8FimeDt
OutputDirectory
=
"Result"
Add
Sequence - search tweets from specific keyword
Multiple Assign
SearchKeyword
=
"#"+vaccine_opt
SearchLang
=
""
Add
Invoke search workflow
"SearchTweets.xaml"
Import Arguments
5
Open Workflow


Python Scope
Do
Load Python Script
filloc
...
Invoke Python Method
Name
"analysis"
...
Get Python Object
PythonObject
out_meth
...
Write Line
Text
fin_out_py


Get access token
2 action(s)
Double click to View
Assign - output file name
OutputFileName
=
String.Format("Tw
Sequence - create datatabl
2 action(s)
Double click to View
Write Line
Text
"Hello_check"
While
Condition
check
Body
Sequence
Write Line
Text
Count_rows_excel.ToString
If
Condition
Count_rows_excel>=3599
Then
Else
Sequence
Sequence



The Python code:

```
import os

import numpy as np

import pandas as pd

from nltk.stem.porter import *

import smtplib

#import matplotlib.pyplot as plt

import neattext as ntx

import nltk

from nltk.stem import WordNetLemmatizer

from nltk.corpus import stopwords

from textblob import TextBlob

from email.message import EmailMessage

import nltk
```

```
nltk.download('vader_lexicon')

from nltk.sentiment.vader import SentimentIntensityAnalyzer

lemmatizer = WordNetLemmatizer()

vader = SentimentIntensityAnalyzer()



def analysis(file,email):



   data=pd.read_excel(r"E:\uipath\SearchTweet\Result\TweetSearch_#"+file+".xls
x",sheet_name = "sheet1")

   def remu(str1):

      out=""

      cnt=0

      for i in str1:

         if(cnt==1):

            out=out+i

         if(i=='>'):

            cnt=1

         elif(i=='<'):

            cnt=0



      return(out[:len(out)-1])

   data1=data.TweetFrom.value_counts()

   data1=data1.rename(lambda x:remu(x))

   data['CreatedAt'] = pd.to_datetime(data['CreatedAt']).dt.date
```

```
   data=data.drop_duplicates('TweetText')

   data.drop(columns={"CreatedAt","UserName","UserId","UserIcon","TweetUrl"},i
nplace=True)

   # Cleaning the data using neattext library

   data['clean_data']=data['TweetText'].apply(ntx.remove_hashtags)

   data['clean_data']=data['clean_data'].apply(ntx.remove_urls)

   data['clean_data']=data['clean_data'].apply(ntx.remove_userhandles)

   data['clean_data']=data['clean_data'].apply(ntx.remove_multiple_spaces)

   data['clean_data']=data['clean_data'].apply(ntx.remove_special_characters)



   stop_words = stopwords.words('english')

   stop_words=stop_words[:147]

   stop_words.remove('not')

   stop_words.remove("don't")



   def stopWords(tweet):

      clean_tweet = tweet

      clean_tweet = " ".join(word for word in clean_tweet.split() if word not
in stop_words)

      return clean_tweet

   data['clean_data'] = data['clean_data'].apply(lambda x: stopWords(x))

   tokenized_tweet = data['clean_data'].apply(lambda x: x.split())

   stemmer = PorterStemmer()
```

```
   # apply stemmer for tokenized_tweet

   tokenized_tweet = tokenized_tweet.apply(lambda x: [stemmer.stem(i) for i in
x])

   tt=list(tokenized_tweet.keys())

   for i in tt:

      tokenized_tweet[i] = ' '.join(tokenized_tweet[i])

   # change df['Tweet'] to tokenized_tweet

   data['clean_data']  = tokenized_tweet

   sentiments = []

   data['clean_data'] = data['clean_data'].astype('string')

   def get_value_counts(col_name, analyzer_name):

      count = pd.DataFrame(data[col_name].value_counts())

      percentage =
pd.DataFrame(data[col_name].value_counts(normalize=True).mul(100))

      value_counts_df = pd.concat([count, percentage], axis = 1)

      value_counts_df = value_counts_df.reset_index()

      value_counts_df.columns = ['sentiment', 'counts', 'percentage']

      value_counts_df.sort_values('sentiment', inplace = True)

      value_counts_df['percentage'] =
value_counts_df['percentage'].apply(lambda x: round(x,2))

      value_counts_df = value_counts_df.reset_index(drop = True)

      value_counts_df['analyzer'] = analyzer_name

      return value_counts_df

   neutral_thresh=0.05
```

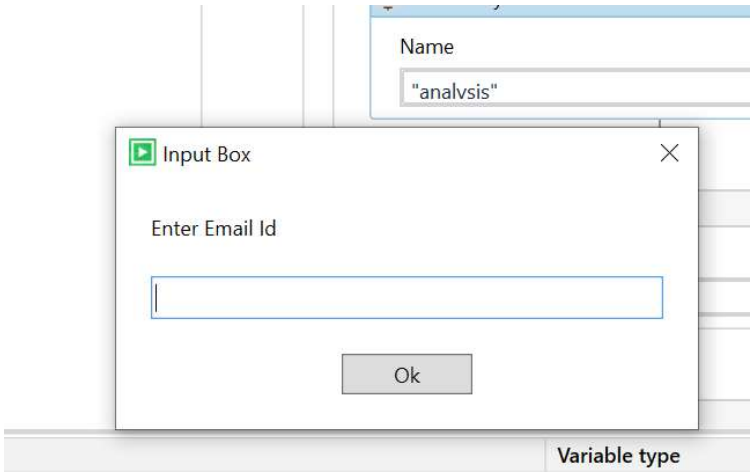
```
   data['textblob_score'] = data['clean_data'].apply(lambda x:
TextBlob(x).sentiment.polarity )

   data['textblob_sentiment'] = data['textblob_score'].apply(lambda c:
'Positive' if c >= neutral_thresh else ('Negative' if c <= -(neutral_thresh)
else 'Neutral'))

   textblob_sentiment_df = get_value_counts('textblob_sentiment','TextBlob')

   for j in data['clean_data']:

      sentiments.append(vader.polarity_scores(j)['compound'])

   data["vader_score"]=sentiments

   data['vader_sentiment'] = data['vader_score'].apply(lambda c: 'Positive' if
c >= neutral_thresh else ('Negative' if c <= -(neutral_thresh) else
'Neutral'))

   vader_sentiment_df = get_value_counts('vader_sentiment','Vader')

   data['composite_score'] =  (data['vader_score']

                               + data['textblob_score']

                               )/2

   data['composite_vote_2'] = data['composite_score'].apply(lambda c:
'Positive' if c >= neutral_thresh else ('Negative' if c <= -(neutral_thresh)
else 'Neutral'))

   composite_sentiment_df_2 = get_value_counts('composite_vote_2','Composite
Sentiment')

   out_str=str(composite_sentiment_df_2['percentage'][0])+"% Negative
"+str(composite_sentiment_df_2['percentage'][1])+"% Neutral
"+str(composite_sentiment_df_2['percentage'][2])+"% Postive "




   msg = EmailMessage()
```

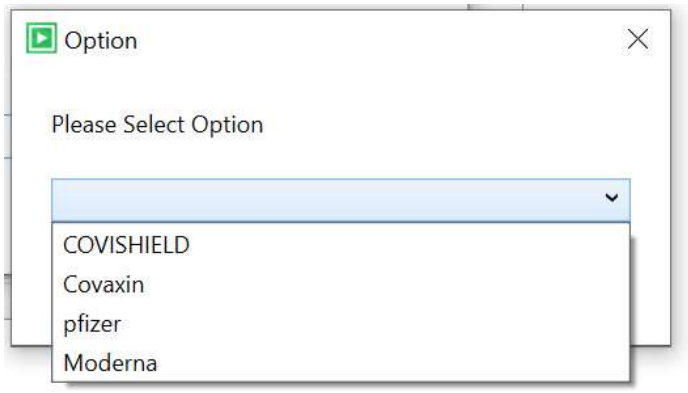
```
    msg.set_content(out_str)
    fromEmail = 'kamisetty.harshith1@gmail.com'
    toEmail = email
    msg['Subject'] = 'Simple Text Message'
    msg['From'] = fromEmail
    msg['To'] = toEmail
    s = smtplib.SMTP('smtp.gmail.com', 587)
    s.starttls()
    s.login(fromEmail, 'Harshith@1089')
    s.send_message(msg)
    s.quit()
    stee="Sucess"
    return stee
analysis("pfizer","121810313024@gitam.in")
```

7.2. TESTING

The flowchart diagram in rpa will be executed in which it collects the tweets and stores in the .xlsx file after which it invokes the python script and sends a result to mail.

Output:





Working:

1) First, Invoke the Rpa Flowchart where it asks the user email id and vaccine name.

2) Based on the selected vaccine, will start collecting the tweets from twitter api with the help of hashtags(#vacccinename)

3) After getting 10,000 rows we will store them in a .xlsx file which will be further used for performing sentimental analysis

4) First step in this analysis is to verify the data consistence like to check null values, column format, and to remove the duplicate

5) After completion of the above step, we will consider only respective coloums which are responsible for performing analysis and we discard the remaning one

6) As in order to perform sentimental analysis on the tweet text we need to clean the text for better result

7) Cleaning text involves removal of #tags, urls, user profiles, punctuation, stop words which donot contribute in the analysis like (a,the,and..etc)

8) After this we again do the Lemmatization for the text in which it switches any kind of a word to its base root mode for this we have used PotterStemer

9) By performing the above steps we will complete the data pre processing after this we will perform the sentimental analysis by using 2 different libraries known as textblob, vader.

10) Where these returns a polarity score range from [-1,1] where we do the ensemble method In which we perform the average of the both scores based upon this score will mark the specific tweet as Postive , Neutral, Negative.

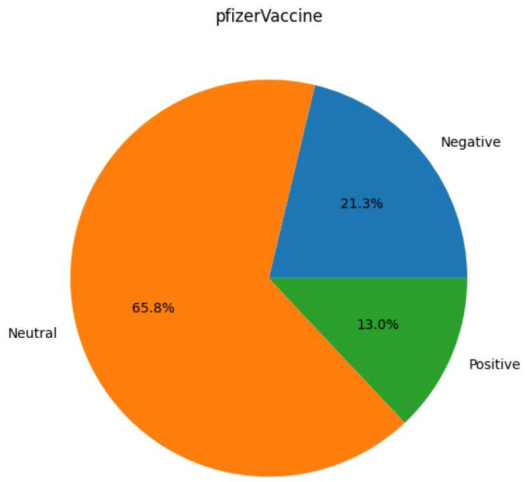
11) After this we will mail the result to the user email id

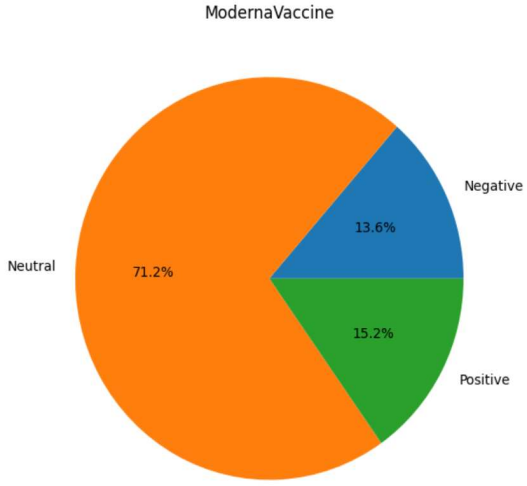
8. RESULTS AND DISCUSSIONS

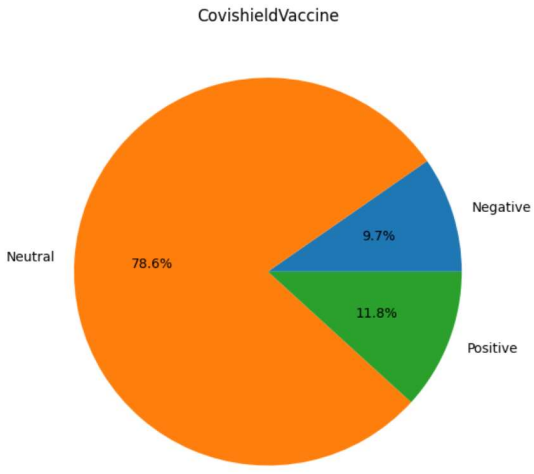
To collect the tweets by using RPA it will take around 15-20 min. Upon collection of tweets it will take 1 -2 mins to perform the analysis and share it via mail. As to perform sentimental analysis we are using 2 pre defined models like textblob, vader because these models are specially trained on social media data.

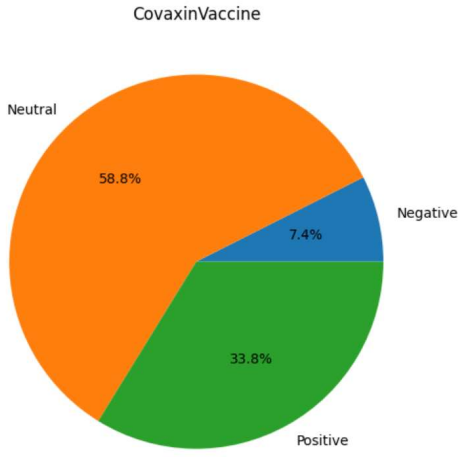
Once the data has been collected with a help of an RPA through twitter API then we perform sentimental analysis using Textblob, Vader which is a sentimental analyzer tools

The output of the model will give you a idea of the opinion of the public on covid vaccine. According to the dataset collected the Moderna vaccines 13.58% Negative ,71.19% Neutral and 15.22% Positive opinion. Where for Pfizer 21.26% Negative, 65.76% Neutral and 12.98% Positive , we have 78.6% Positive ,9.7% negative and 11.8% Postive, where for covaxin 33.8% Positive, 58.8% Neutral and 7.4% Negative.









9. CONCLUSION AND FUTURE SCOPE

While it is important for the Indian government to actively encourage its citizens to have vaccine, it is also important to help the citizens understand the important of the vaccination program. The best way to educate citizens regarding the positive aspect of the vaccination program is by knowing the public opinion on different vaccines where Indian citizens have voiced in their social media post about the COVID-19 vaccines by analyzing these tweets we can get the sentiments about particular vaccine.

By this reviews a user can decide to choose which type of vaccine he/she intended to take.

10. REFERENCES

1. Naw Safrin Sattar, Shaikh Arifuzzaman, (2015), COVID-19 Vaccination Awareness and Aftermath: Public Sentiment Analysis on Twitter Data and Vaccinated Population Prediction in the USA
2. Loria, S. textblob Documentation. Release 0.16. 26 April 2020. Available online: https://buildmedia.readthedocs.org/media/pdf/textblob/latest/textblob.pdf
3. Hutto, C.; Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the International AAAI Conference on Web and Social Media, Ann Arbor, MI, USA, 1–4 June 2014; Volume 8