CAPSTONE PROJECT WORK REPORT

Phase II

Omicron Sentiment Analysis

SANJEETH M

A report submitted in part fulfilment of the degree of

B.Sc. in Computer Science with Data Analytics

Supervisor: **Mrs.P.JAYAPRIYA**, MCA,M.E.,(Ph.D.)
Associate professor, Dept. of CS with DA

Department of Computer Science with Data Analytics

KPR College of Arts Science and Research
(Affiliated to Bharathiar University, Coimbatore)
Avinashi Road, Arasur, Coimbatore – 641 407

NOVEMBER 2022

CAPSTONE PROJECT WORK REPORT

Phase II

Omicron Sentiment Analysis

Bonafide Work Done by

SANJEETH M

REG. NO. 2028B0044

KPR College of Arts Science and Research
Learn Beyond
Avinashi Road, Arasur, Coimbatore.

Dissertation submitted in partial fulfillment of the requirements for the award of Bharathiar University, Coimbatore-46.

Signature of the Guide **Signature of the HOD**

[Mrs.P.Jayapriya]

Submitted for the Viva-Voce Examination held on ____________________

Internal Examiner **External Examiner**

CONTENTS

ACKNOWLEDGEMENT

In the accomplishment of completion of my Capstone Project Work Phase - II on **Omicron Sentiment Analysis using Python** I would like to convey my special gratitude to **Dr. S. Balusamy, Principal of KPR College of Arts Science and Research** and **Mrs.P.Jayapriya, Associate Professor, Department of Computer Science with Data Analytics**. Your valuable guidance and suggestions helped me in phase - II of the completion of this project. I will always be thankful to you in this regard. I am ensuring that this project was finished by me and not copied.

Student Signature

Place:

Date:

KPR COLLEGE OF ARTS SCIENCE AND RESEARCH

(Affiliated to Bharathiar University, Coimbatore)

Avinashi Road, Arasur, Coimbatore – 641 407

ABOUT THE COLLEGE

KPR College of Arts Science and Research is the latest addition to the KPR fleet. The College is located in a picturesque campus of about 11. Acres. The College is run by KPR charities under the leadership of our Chairman Dr. K.P. Ramasamy. The KPR Group is one of the largest industrial conglomerate in the country with interest in Textiles, Sugar, Wind Turbines, Automobiles and Education. The College was established in the year 2019 with a vision of providing top class education and life skills to students and thereby serve the nation and beyond. KPRCAS today offers 12 UG programmes in Management, Commerce and Computer Science streams. The Students of KPRCAS undergo intense training not only in the syllabus and curriculum of the affiliating University but are also trained in various areas. So that they emerge as industry ready graduates to meet the varying demands of the competing industries. Character building and Leadership qualities are inculcated into the students to make them responsible citizens focusing on the development of society and nation. A plethora of Clubs and Events encouraged the students to take part in sports and other cultural activities. KPRCAS offers three years undergraduate courses, which are exclusively for Business, Commerce and Computer Science Stream. The students are equipped with skills and knowledge needed to take up various leadership positions and to develop the society. Beyond Book Teaching help them to be professionals. KPRCAS emphasis on making the students academically brilliant, and also prepare them for the real corporate world. The learning curve begins here for the students of KPRCAS.

ABOUT THE DEPARTMENT

Bachelor of Computer Science with Data Analytics (B.Sc. (CS with DA)) was established in the year 2020. Data Analytics helps to raise the quality of data in the entire business system. The goal of data analytics is to construct the means for extracting business-focused insights from data This requires an understanding of how value and information flows in a business, and the ability to use that understanding to identify business opportunities. The primary aim of a data analyst is to increase efficiency and improve performance by discovering patterns in data. Data analysts exist at the intersection of information technology, statistics and business. They combine these fields in order to help businesses and organizations succeed. The students get exposed to Big Data, Business Intelligence, Data Mining, Data Visualization, Advanced Excel, Predictive Analytics and R Programming.

SYNOPSIS

Twitter is a miniature writing for a blog site which gives phase to individuals to share as well as communicate their perspectives about point, activities, items plus other medicinal harms. Tweets can be arranged keen on assorted classes reliant on their significance through the tip looked. For genuine effecting of this structure python through NLP plus twitter informational compilation be used. In this project we are concerning feelings exploration in twitter tweet for omicron datasets to arrange the survey of all consumers whether it is positive, negative or impartial.

The WHO designated variant of the coronavirus, B.1.1.529, as a variant of concern which has been named Omicron. Right after that, we saw an outbreak of tweets about the Omicron variant on Twitter. In this project, walk through the task of Omicron Sentiment Analysis using Python. Sentiment analysis is the task of natural language processing where we detect a positive, negative or impartial sentiment from a piece of text. Sentiment analysis is used by companies to analyze the opinions of customers about their products or services so that they can use the positive sentiments to market their products or services and the negative sentiments to improve the quality of their products or services, same strategy applied in our project.

CHAPTER 1

1.INTRODUCTION

1.1 Sentiment Analysis

Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material, and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations. Analysis of social media streams is usually restricted to just basic sentiment analysis and count based metrics. This is akin to just scratching the surface and missing out on those high value insights that are waiting to be discovered. With the recent advances in deep learning, the ability of algorithms to analyse text has improved considerably. Creative use of advanced artificial intelligence techniques can be an effective tool for doing in-depth research. We believe it is important to classify incoming customer conversation about a brand based on following lines:

- Key aspects of a brand’s product and service that customers care about.
- Users’ underlying intentions and reactions concerning those aspects.

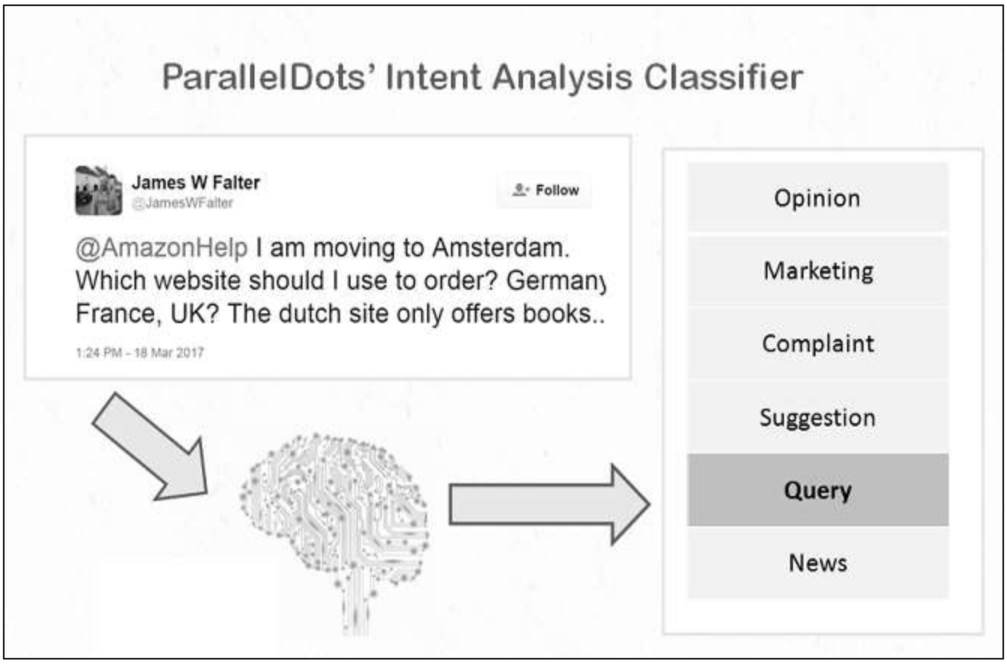
These basic concepts when used in combination, become a very important tool for analysing millions of brand conversations with human level accuracy.

1.2 Text Classifier

Sentiment Analysis is the most common text classification tool that analyses an incoming message and tells whether the underlying sentiment is positive, negative our neutral.

1.3 Intent Analysis

Intent analysis steps up the game by analysing the user’s intention behind a message and identifying whether it relates an opinion, news, marketing, complaint, suggestion, appreciation or query.



@AmazonHelp I am moving to Amsterdam. Which website should I use to order? German) France, UK? The dutch site only offers books..

1.3.1 Analysing intent of textual data

1.4 Sentiment Library

The way your brain remembers the descriptive words you encounter over your lifetime and their relative "sentiment weight", a basic sentiment analysis system draws on a sentiment library to understand the sentiment-bearing phrases it encounters. Sentiment libraries are very large collections of adjectives (good, wonderful, awful, horrible) and phrases (good game, wonderful story, awful performance, horrible show) that have been hand-scored by human coders. This manual sentiment scoring is a tricky process, because everyone involved needs to reach some agreement on how strong or weak each score should be relative to the other scores. If one person gives "bad" a sentiment score of -0.5, but another person gives "awful" the same score, your sentiment analysis system will conclude that that both words are equally negative.

A multilingual sentiment analysis engine must maintain unique libraries for each language it supports. And each of these libraries must be maintained constantly: scores tweaked, new phrases added, irrelevant phrases removed.

1.5 Rules-Based Sentiment Analysis Systems

The sentiment libraries are prepared, software engineers write a series of guidelines (“rules”) to help the computer evaluate the sentiment expressed towards a particular entity (noun or pronoun) based on its nearness to known positive and negative words (adjectives and adverbs). To continue our baseball example, an engineer might create search rules that look like:

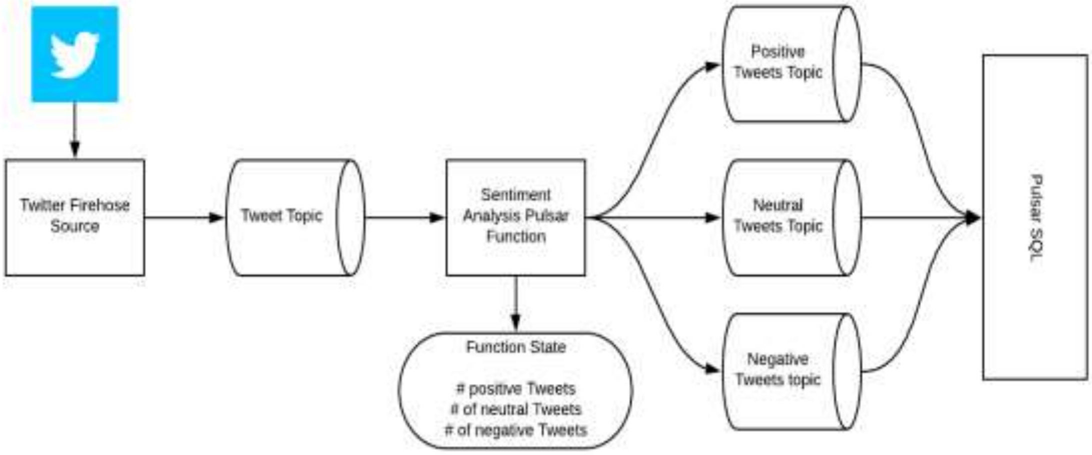
(pitching) near (good, wonderful, spectacular)

(pitching) near (bad, horrible, awful)

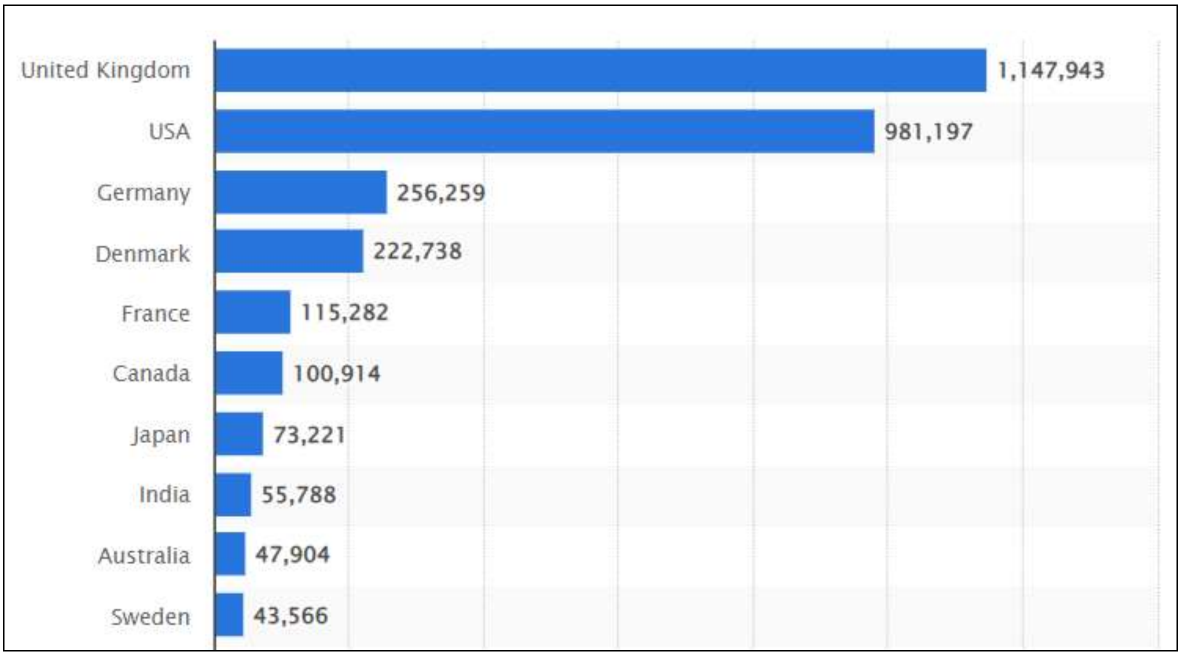
These queries return a “hit count” representing how many times the word “pitching” appears near each adjective. The system then combines these hit counts using a complex mathematical operation called a “log odds ratio”. The outcome is a numerical sentiment score for each phrase, usually on a scale of -1 (very negative) to +1 (very positive).

1.6 Omicron Variant

This research presents the findings of an exploratory study on the continuously generating Big Data on Twitter related to the sharing of information, news, views, opinions, ideas, knowledge, feedback, and experiences about the COVID-19 pandemic, with a specific focus on the Omicron variant, which is the globally dominant variant of SARS-CoV-2 at this time. A total of 12028 tweets about the Omicron variant were studied, and the specific characteristics of tweets that were analyzed include - sentiment, language, source, type, and embedded URLs. The findings of this study are manifold. First, from sentiment analysis, it was observed that 50.5% of tweets had the ‘neutral’ emotion. The other emotions - ‘bad’, ‘good’, ‘terrible’, and ‘great’ were found in 15.6%, 14.0%, 12.5%, and 7.5% of the tweets, respectively. Second, the findings of language interpretation showed that 65.9% of the tweets were posted in English. It was followed by Spanish or Castillian, French, Italian, Japanese, and other languages, which were found in 10.5%, 5.1%, 3.3%, 2.5%, and &lt;2% of the tweets, respectively. Third, the findings from source tracking showed that “Twitter for Android” was associated with 35.2% of tweets. It was followed by “Twitter Web App”, “Twitter for iPhone”, “Twitter for iPad”, “Tweet Deck”, and all other sources that accounted for 29.2%, 25.8%, 3.8%, 1.6%, and &lt;1% of the tweets, respectively. Fourth, studying the type of tweets revealed that retweets accounted for 60.8% of the tweets, it was followed by original tweets and replies that accounted for 19.8% and 19.4% of the tweets, respectively. Finally, in terms of embedded URL analysis, the most common URLs embedded in the tweets were found to be twitter.com, which was followed by biorxiv.org, nature.com, wapo.st, nzherald.co.nz, recvprofits.com, science.org, and other URLs.



1.6.1 Sentiment Analysis of Tweets



1.6.2. SARS-CoV-2 Omicron variant cases by country or territory

CHAPTER 2

2. SYSTEM SPECIFICATION

2.1. Hardware Configuration

Operating System	**Self-Hosted Technical Requirement**	**Cloud Technical Requirement**
Windows	Windows 8.1+	Windows 8.1+
Mac	Mac OS 10.14+	Mac OS 10.14+
Linux	Ubuntu LTS releases 18.04 or later	Ubuntu LTS releases 18.04 or later
RAM	8 GB	
HDD	1 TB	
Processor	64-bit, four-core, 2.5 GHz minimum per core (If your dataset size is significantly larger than the medium dataset, we recommend 8 cores.)	
Mouse	Dell MS116 1000DPI USB Wired Optical Mouse	
Keyboard	Dell KB522 Business Keyboard-Black	
Monitor	Dell 24 Monitor-S2421HN in-Plane Switching (IPS)	

2.3. Software ware Configuration

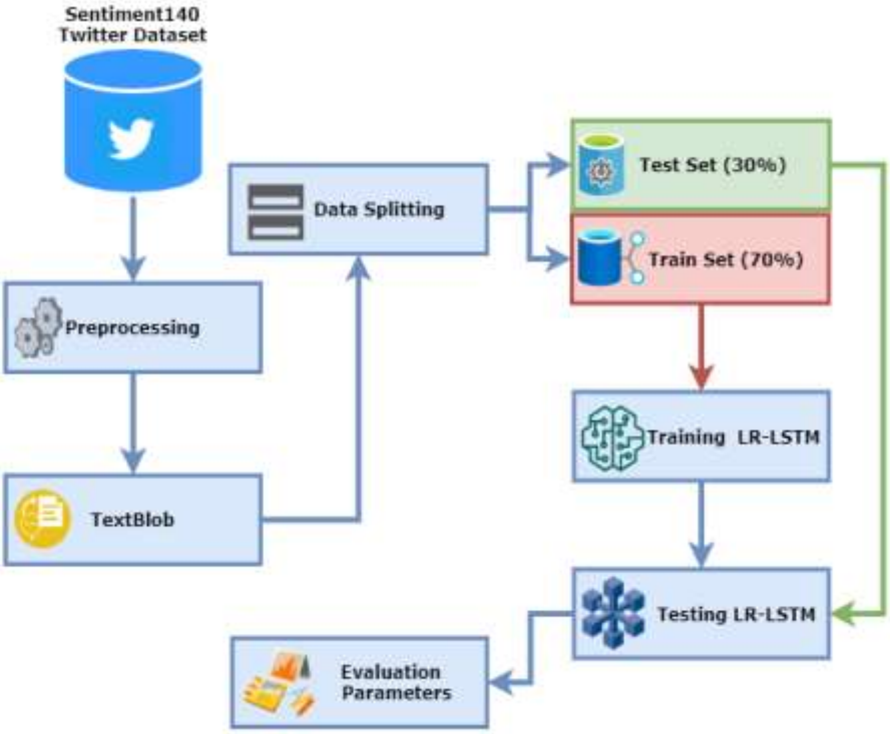
IDE	Anaconda
Language Support	Python 3.9
Platform	Jupyter Notebook
Browser	Google Chrome Version 101.0.4951.67
Database	MySQL 8.0.29

CHAPTER 3

3. SYSTEM STUDY

3.1 Existing System

Several methods are available in the literature, that use base classifiers for Twitter. Presents a survey of Sentiment Analysis algorithms and applications determined sentiments by using emoticons and hashtag integrate lexicon and learning based techniques. They used lexicons and POS as linguistic resources. Proposed an efficient ensemble classifier using a multi-objective differential evolution algorithm. They compared weighted and unweighted voting schemes. They made no attempt to perform any data pre-processing. Compared feature hashing and bag-of-words for feature representation. They proposed ensemble classifier based on ma jority vote for Twitter sentiment analysis. Used feature hashing as feature representation technique and logistic regression as a base learner. Linguistic processing is not fully covered in this paper. The result shows that ensemble classifier performed well. Used N-gram, lexicon, POS and Sentiwordnet as feature set.



3.1.1 Tweet sentiment classification approach using ensemble classifier.

SVMs and Conditional Random Fields are used as base learner. Their ensemble combination of orthogonal methods leads to more accurate classifiers. Developed an ensemble technique which used dataset, feature set and bootstrap aggregation learners. They proposed an algorithm that would select the most appropriate classifier among all the base classifiers. Proposed an ensemble classifier which is trained on features like lexical to determine the polarity of each individual phrase within each tweet. The sentiment of a specific phrase may not be same as the sentiment of the whole tweet.

Some of the other popular use cases include improved search, improved tweet contents, and predicting election outcomes. Reviewing studies catering to these use cases is an important tool for identifying the techniques, which can help improve the impact and effectiveness of the recommendation system. Guo and Lease proposed a novel ranking model, for enriching the search functionality on Twitter, with personalization and content analysis. Clark and Araki introduced a text normalization technique to categorize errors and informal language used on social media into different groups, followed by natural language processing techniques to correct common phonetic and slang mistakes. On the contrary, Laniado and Peter applied hashtags on Twitter and demonstrated mappings of fifty percent hashtags to entities in freebase.

The system was categorized into four dimensions:

frequency, specificity, consistency, and stability to assess hashtags as strong identifiers. Losch and Müller proposed a method to associate hashtags with encyclopedia entities. Their system used Wikipedia entities as a description of hashtags in microblogging service to understand the actual context of hashtags. Tumasjan nalyses Twitter as a source of predicting elections. They used the context of the German federal election to investigate whether Twitter is used as a forum for political deliberation. They used LIWC 2007, a text analysis software, which uses a psychometrically validated dictionary for identifying and assessing the emotional, cognitive, and structural components of given text samples. The authors used 12 dimensions including past and future orientation, positive and negative emotion, sadness, anxiety, anger, tentativeness, certainty, work, achievement, and money to extract political sentiments from this data.

3.2 Drawbacks

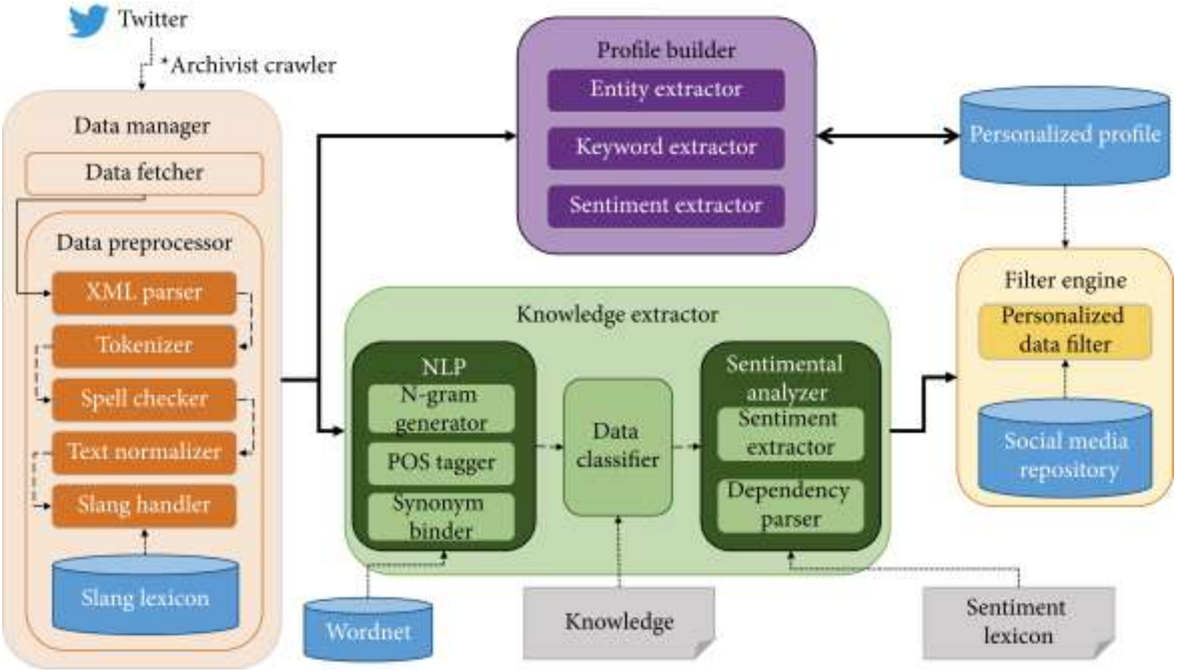
Due to Twitter’s accessible and easy-to-use API, it has become common to use the social media platform as a data source for data science projects. It’s easy to search for all the occurrences of a hashtag, for example, and convince yourself that you’re tapping in to the worldwide conversation on a given issue. But while Twitter is an excellent source of data overall, we should be careful about the kinds of conclusions we draw from Twitter data. In this article, we want to outline a few of the challenges that we can face when we decide to use Twitter as a data source. Some of these are technical issues that can be overcome, others are endemic problems that we need to bear in mind when interpreting our results.

Tweets by bots can be hard to detect, especially with small sample sizes. Yet they can skew your data, just as they are occasionally intended to purposively skew the conversation. Indeed, bots tend to tweet more often and to use key terms more frequently than normal human beings because they are focused in on specific elements of a conversation. This means that they create more noise in the data than a normal human being might.

3.3 Proposed System

Twitter is a popular social media platform that enables users to post short texts, images, and videos of personal and/or collaborative nature. This data provides a unique insight into the user’s personality. Of particular interest to our research work, are the user’s interests and emotions, which are used by our proposed system to build a user profile and then provide personalized data/services to similar users. The dataset that we using for the task of Omicron sentiment analysis is collected from tweet download, which was initially collected from Twitter when people were sharing their opinions about the Omicron variant.

3.4 System Architecture



3.4.1. TSA – System Architecture

To classify tweets and extract topic level sentiments, the system nalyses tweets using domain-specific seed words, opinion words, n-gram generator, POS tagger, synonym binder, and dependency parser. Seed words and opinion words are enriched by synonyms to increase accuracy of classification.

3.5 Sentiment Analysis on Twitter Data

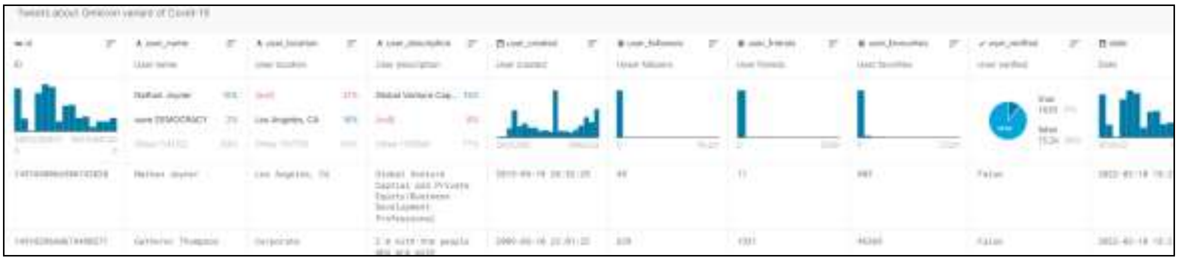
Performing sentiment analysis on Twitter data involves five steps:

1. Gather relevant Twitter data
2. Clean your data using pre-processing techniques
3. Create a sentiment analysis machine learning model
4. Analyze your Twitter data using your sentiment analysis model
5. Visualize the results of your Twitter sentiment analysis

3.6 Prepare Your Data

These tweets are collected using Twitter API and a Python script. A query for this high-frequency hashtag (#Omicro) is run on a daily basis for a certain time period, to collect a larger number of tweets samples. Once you have gathered the tweets you need for your sentiment analysis, you'll need to prepare your data. Social media data is unstructured and needs to be cleaned before using it to train a sentiment analysis model – good quality data will lead to more accurate results.

Preprocessing a Twitter dataset involves a series of tasks like removing all types of irrelevant information like emojis, special characters, and extra blank spaces. It can also involve making format improvements, delete duplicate tweets, or tweets that are shorter than three characters.



Tweets about Omicron variant of Covid-19

id	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date
ID	User name	User location	User description	User created	User followers	User friends	User favourites	User verified	Date
[illegible]	Nathan Joyner	Los Angeles, CA	Global Venture Capital and Private Equity/Business Development Professional	[illegible]	44	11	[illegible]	False	2022-01-19 [illegible]
[illegible]	Catherine Thompson	Corporate	I am with the people who are with	[illegible]	[illegible]	[illegible]	[illegible]	False	2022-01-19 [illegible]

3.6.1 Tweets about Omicron variant of Covid-19

We tried to build a sentiment analysis system by studying and implementing algorithms of machine learning. We implemented Naïve Bayes and Maximum Entropy algorithms. Baseline model performed the worst with no doubt as it had least number of features. The modular system we've built can easy be scaled for new algorithms be it in Machine Learning, Deep learning or Natural Language Processing. Sentiment analysis system is an active field of research and we can still further improve our system by working more on the algorithms, trying out different things in preprocessing and checking which ones get the best precision metrics.

3.7 Features

The use of this information can be applied to make wiser decisions related to the use of resources, to make improvements in Medical Field, providing better Medicine /services, and ultimately to improve the patient lifestyle. An example of this application is the impact of tracking people's feelings on products, services and events, which allow enterprise managers to have knowledge and parameters to decision-making. Another example is city council administrators that could have the opportunity for improving the services offered to citizens and for addressing challenges of development and sustainability more efficiently based on what people feel. Social media is the current environment for data collection and analysis of sentiments of people. People can share and comment on everything, from personal thoughts to common events or topics in society. The access to social media also can provide more information in the form of hidden metadata. For instance, Operating System language, device type, capture time and geographical location.

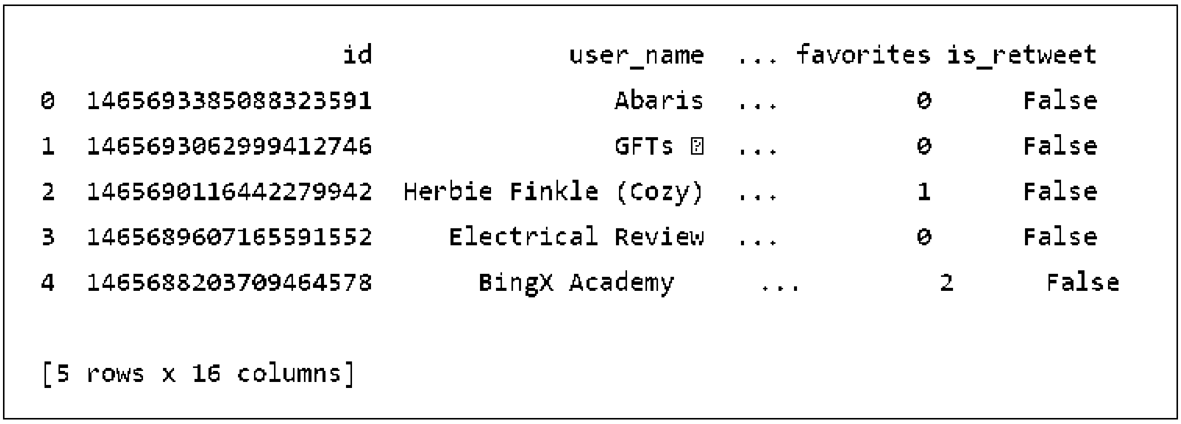
CHAPTER 4

4. SYSTEM DESIGN

4.1 Sentiment Analysis using Python

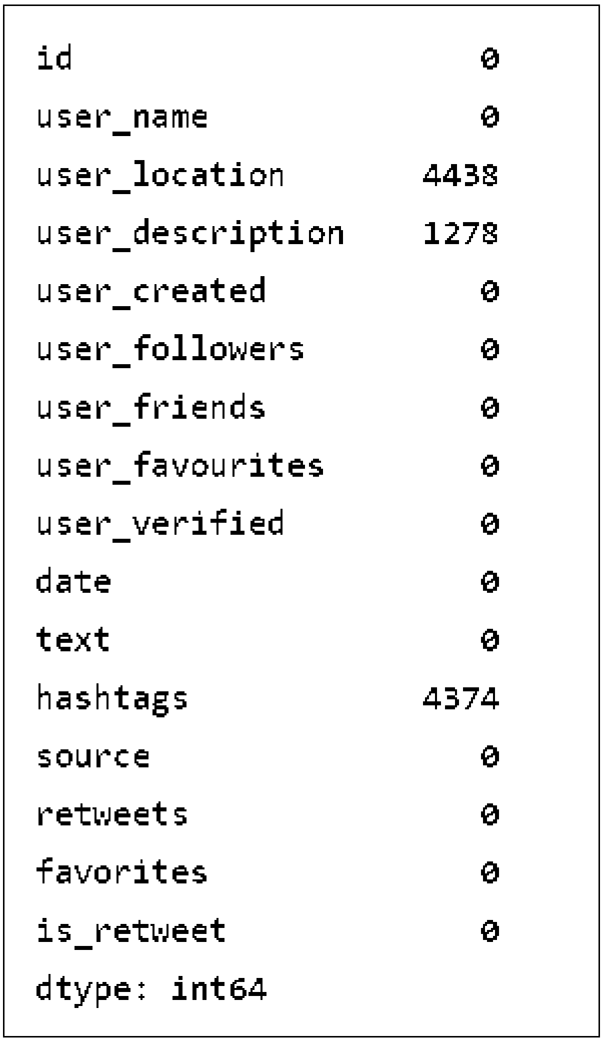
Start the task of Omicron sentiment analysis by importing the necessary Python libraries and the dataset.

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from wordcloud import WordCloud, STOPWORDS,
ImageColorGenerator

data = pd.read_csv("omicron.csv")
print(data.head())
```



```
                    id            user_name  ... favorites is_retweet
0  1465693385088323591               Abaris  ...         0      False
1  1465693062999412746               GFTs ⍰  ...         0      False
2  1465690116442279942  Herbie Finkle (Cozy)  ...         1      False
3  1465689607165591552    Electrical Review  ...         0      False
4  1465688203709464578        BingX Academy    ...         2      False

[5 rows x 16 columns]
```

```
print(data.isnull().sum())
```



```
id                     0
user_name              0
user_location       4438
user_description    1278
user_created           0
user_followers         0
user_friends           0
user_favourites        0
user_verified          0
date                   0
text                   0
hashtags            4374
source                 0
retweets               0
favorites              0
is_retweet             0
dtype: int64
```

The dataset contains null values in three columns that contains textual data, we will remove all the rows containing the null values:

data = data.dropna()

4.2 Sentiment Analysis of Omicron Variant

The text column in the dataset contains the tweets done by people to share their opinions about the Omicron variant. To move further, we need to clean and prepare this column for the task of sentiment analysis.

import nltk

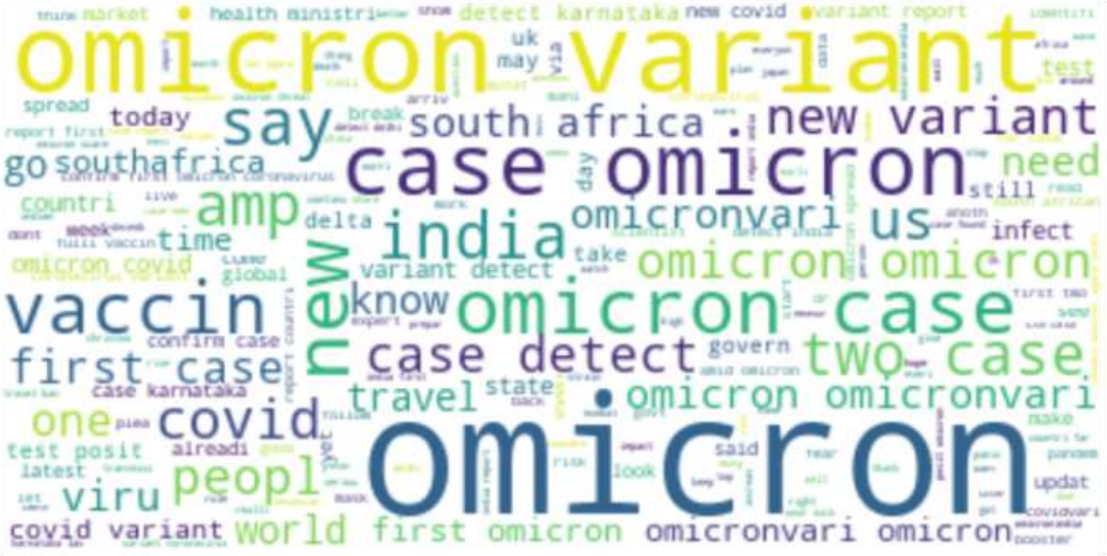
import re

```
nltk.download('stopwords')

stemmer = nltk.SnowballStemmer("english")

from nltk.corpus import stopwords

import string

stopword=set(stopwords.words('english'))

def clean(text):

    text = str(text).lower()

    text = re.sub('\[.*?\]', '', text)

    text = re.sub('https?://\S+|www\.\S+', '', text)

    text = re.sub('<.*?>+', '', text)

    text = re.sub('[%s]' % re.escape(string.punctuation), '', text)

    text = re.sub('\n', '', text)

    text = re.sub('\w*\d\w*', '', text)

    text = [word for word in text.split(' ') if word not in stopword]

    text=" ".join(text)

    text = [stemmer.stem(word) for word in text.split(' ')]

    text=" ".join(text)

    return text
  data["text"] = data["text"].apply(clean)
```

As we have cleaned the text column, now let’s have a look at the word cloud of the text column to look at the most number of words used by the people on their tweets:

```
text = " ".join(i for i in data.text)

stopwords = set(STOPWORDS)
```

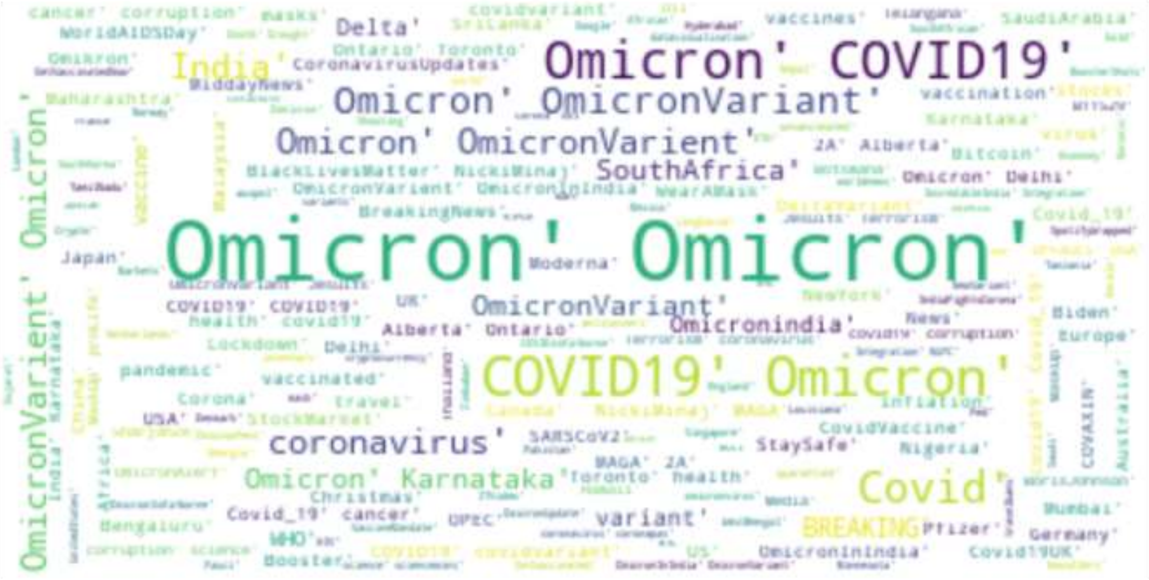
```
wordcloud = WordCloud(stopwords=stopwords,
background_color="white").generate(text)

plt.figure( figsize=(15,10))

plt.imshow(wordcloud, interpolation='bilinear')

plt.axis("off")

plt.show()
```



4.2.1. word cloud of the hashtags

The word cloud of the hashtags column to look at the most number of hashtags used by the people on their tweets:

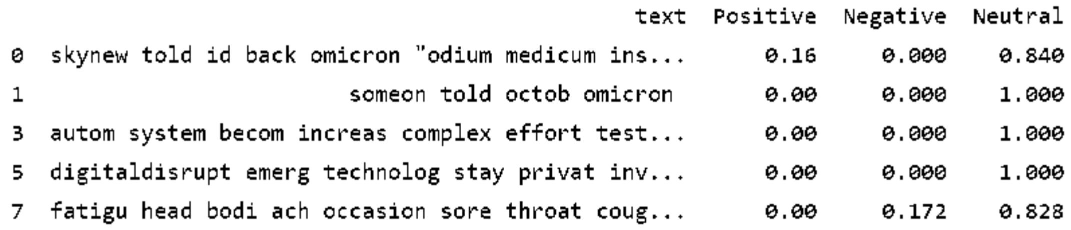
```
text = " ".join(i for i in data.hashtags)

stopwords = set(STOPWORDS)

wordcloud = WordCloud(stopwords=stopwords,
background_color="white").generate(text)

plt.figure( figsize=(15,10))

plt.imshow(wordcloud, interpolation='bilinear')

plt.axis("off")
plt.show()
```



4.2.2. word cloud of the hashtags -2

We will calculate the sentiment scores of the tweets about the Omicron variant. Here we will add three more columns in this dataset as Positive, Negative, and Neutral by calculating the sentiment scores of the text column.

```
nltk.download('vader_lexicon')

sentiments = SentimentIntensityAnalyzer()

data["Positive"] = [sentiments.polarity_scores(i)["pos"] for i in data["text"]]

data["Negative"] = [sentiments.polarity_scores(i)["neg"] for i in
data["text"]]

data["Neutral"] = [sentiments.polarity_scores(i)["neu"] for i in data["text"]]

data = data[["text", "Positive", "Negative", "Neutral"]]

print(data.head())
```



```
                                                text  Positive  Negative  Neutral
0  skynew told id back omicron "odium medicum ins...      0.16     0.000    0.840
1                           someon told octob omicron      0.00     0.000    1.000
3  autom system becom increas complex effort test...      0.00     0.000    1.000
5  digitaldisrupt emerg technolog stay privat inv...      0.00     0.000    1.000
7  fatigu head bodi ach occasion sore throat coug...      0.00     0.172    0.828
```

4.3 People reacted about the Omicron variant

```
x = sum(data["Positive"])
y = sum(data["Negative"])
z = sum(data["Neutral"])
def sentiment_score(a, b, c):
if (a>b) and (a>c):
print("Positive 😊 ")
elif (b>a) and (b>c):
print("Negative 😬 ")
else:
print("Neutral 🙂 ")
sentiment_score(x, y, z)
```

Output:

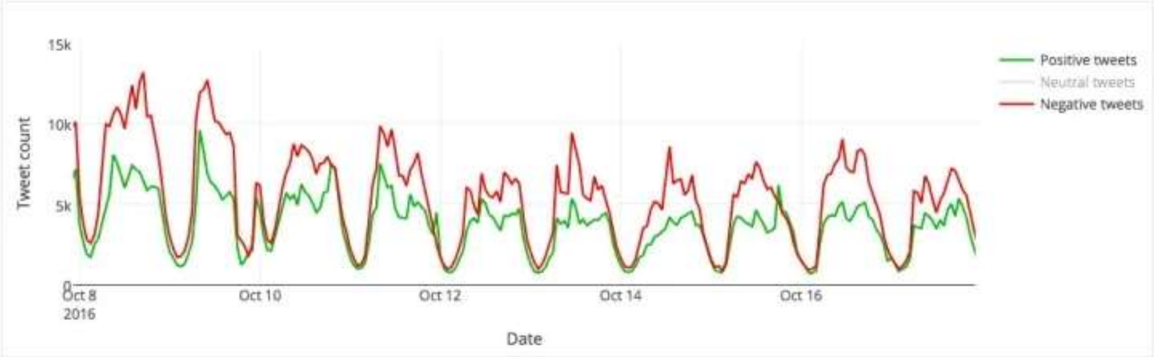
```
Neutral 🙂
```

4.4 Database Design

USER_NAME	USER_LOCATION	USER_DESCRIPTION	USER_CREATED	USER_FOLLOWERS	USER_FRIENDS	USER_FAVOURITES
Tom Basile ðŸ‡ºðŸ‡¸	New York, NY	Husband, Father, Columnist & Commentator. Author of Tough Sell: Fighting the Media War in Iraq. Bush Admin Alum. Newsmax Contributor. Fmr Exec Dir NYSGOP	4/16/2009 20:06	2253	1677	24
Time4fisticuffs	Pewee Valley, KY	#Christian #Catholic #Conservative #Reagan #Republican #Capitalist; Sports lover - #BBN #Cincinnati #Reds #Bengals #Trump2020	2/28/2009 18:57	9275	9525	7254
ethel mertz	Stuck in the Middle	#Browns #Indians #ClevelandProud #[]_[] #Cavs #Resist	3/7/2019 1:45	197	987	1488
DIPR-J&K	Jammu and Kashmir	ðŸ–Šï¸□Official Twitter handle of Department of Information and Public Relations, Govt of Jammu & Kashmir	2/12/2017 6:45	101009	168	101
ðŸŽ¹ Franz Schubert	Ð□Ð¾Ð²Ð¾Ñ€Ð¾Ì□Ñ□Ñ□Ð¸Ñ□	ðŸŽ¼ #Ð□Ð¾Ð²Ð¾Ñ€Ð¾Ì□Ñ□Ñ□Ð¸Ñ□ #Novorossiya #Ð¾Ñ□Ñ‚Ð°Ð²Ð°Ð¹Ñ□Ñ□Ð´Ð¾Ð¼Ð° #STAYatHOME Polymath, composer, English.	3/19/2018 16:29	1180	1071	1287

4.5 Data Visualization

Data visualization tools help explain sentiment analysis results in a simple and effective way. Twitter sentiment analysis provides many exciting opportunities. Being able to analyze tweets in real-time, and determine the sentiment that underlies each message, adds a new dimension to social media monitoring.



So most of the opinions were Neutral, which means that people were sharing information about the Omicron variant instead of sharing any positive or negative opinions.

CHAPTER 5

5. CONCLUSION

The Omicron variant of coronavirus a new variant of coronavirus that has been designated as the variant of concern by the World Health Organization. By integrating our proposed system with Twitter, the user would be able to get precisely classified and personalized data with sentiment value. Moreover, this tweet data is useful for clustering, trend analysis, and recommendations as well. In future, we are planning to integrate user information from other social media and user activities log to find interesting patterns and use them in personalized recommender systems.

5.1 BIBLIOGRAPHY

1. Sentiment Analysis using Pythonin 2021 by Aman Kharwal, https://thecleverprogrammer.com

2. Digital in 2017: Global Overview, We Are Social, 2017, https://wearesocial.com/sg/blog/2017/01/digital-in-2017-global-overview.

3. R. Batool, W. A. Khan, M. Hussain et al., "Towards personalized health profiling in social network," in *Proceedings of the 6th International Conference on New Trends in Information Science and Service Science and Data Mining (ISSDM)*, IEEE, Taipei, Taiwan, October 2012.View at: Google Scholar

4. Company Info, Facebook Newsroom, 2019, http://newsroom.fb.com/company-info/.

5. Most Popular Social Networks Worldwide as of January 2019, Ranked by Number of Active Users (in Millions), 2019, https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

6. Q. You, S. Bhatia, and J. Luo, "A picture tells a thousand words-About you! User interest profiling from user generated visual content," *Signal Processing*, vol. 124, pp. 45–53, 2016.View at: Publisher Site | Google Scholar

7. F. Persia and D. D'Auria, "A survey of online social networks: challenges and opportunities," in *Proceedings of the 2017 IEEE International Conference on Information Reuse and Integration, IRI 2017*, pp. 614–620, San Diego, CA, USA, August 2017.View at: Google Scholar

8. A. V. Lakshmi, S. B. R. Kumar, P. J. Charles et al., "Survey paper on mobile social networks," *International Research Journal of Engineering and Technology*, vol. 2, no. 6, pp. 637–641, 2015.View at: Google Scholar

9. A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic," *PLoS One*, vol. 6, no. 5, Article ID e19467, 2011.View at: Publisher Site | Google Scholar

10. A. Weiler, M. Grossniklaus, M. H. Scholl et al., "Survey and experimental analysis of event detection techniques for twitter," *The Computer Journal*, vol. 60, no. 3, pp. 329–346, 2017.View at: Google Scholar

11. A. Crisci, V. Grasso, P. Nesi et al., "Predicting TV programme audience by using Twitter based metrics," *Multimedia Tools and Applications*, vol. 77, no. 3, pp. 12203–12232, 2018.View at: Google Scholar

12. S. J. McConnell, "Twitter and the 2016 U. S. presidential campaign: a rhetorical analysis of a rhetorical analysis of tweets and media coverage by Stephen J. Mcconnell," A thesis Submitted in Partial Fulfillment of the Degree of Master of Science in Professional Writing December 2015 New York University School of Professional Studies, New York, NY, USA, 2016.View at: Google Scholar

13. E. Mohammadi, M. Thelwall, M. Kwasny, and K. L. Holmes, "Academic information on twitter: a user survey," *PLoS One*, vol. 13, no. 5, Article ID e0197265, 2018.View at: Publisher Site | Google Scholar

14. H. S. Ibrahim, S. M. Abdou, and M. Gheith, "Sentiment analysis for modern standard Arabic and colloquial," 2015, https://arxiv.org/abs/1505.03105.View at: Google Scholar

15. B. Al-Jenaibi, "The twitter revolution in the gulf countries," *Journal of Creative Communications*, vol. 11, no. 1, pp. 61–83, 2016.View at: Publisher Site | Google Scholar

16. L. M. Yonker, S. Zan, C. V. Scirica, K. Jethwani, and T. B. Kinane, ""Friending" teens: systematic review of social media in adolescent and young adult health care," *Journal of Medical Internet Research*, vol. 17, no. 1, p. e4, 2015.View at: Publisher Site | Google Scholar

17. Networking Health: Prescriptions for the Internet, 2017, http://www.ncbi.nlm.nih.gov/books/NBK44714/.

18. S. Fox, "Health topics," 2017, http://www.pewinternet.org/2011/02/01/health-topics-2/.View at: Google Scholar

19. M. Ybarra and M. Suman, “Reasons, assessments and actions taken: sex and age differences in uses of Internet health information,” *Health Education Research*, vol. 23, no. 3, pp. 512–521, 2008.View at: Publisher Site | Google Scholar

20. S. S. Tan and N. Goonawardene, “Internet health information seeking and the patient-physician relationship: a systematic review corresponding author,” *Journal of Medical Internet Research*, vol. 19, no. 1, p. e9, 2017.View at: Publisher Site | Google Scholar

21. E. Basch, A. M. Deal, M. G. Kris et al., “Symptom monitoring with patient-reported outcomes during routine cancer treatment: a randomized controlled trial,” *Journal of Clinical Oncology*, vol. 34, no. 6, pp. 557–565, 2019.View at: Publisher Site | Google Scholar

22. P. H. Keckley and M. Hoffmann, *Social Networks in Health Care: Communication, Collaboration and Insights*, Deloitte Center for Health Solutions, New York, NY, USA, 2010.

23. Patientslikeme, 2017, https://www.patientslikeme.com/.

24. A. Sarker, R. Ginn, A. Nikfarjam et al., “Utilizing social media data for pharmacovigilance: a review,” *Journal of Biomedical Informatics*, vol. 54, no. 1, pp. 202–212.View at: Google Scholar

25. V. Ehrenstein, H. Nielsen, A. B. Pedersen, S. P. Johnsen, and L. Pedersen, “Clinical epidemiology in the era of big data: new opportunities, familiar challenges,” *Clinical Epidemiology*, vol. 9, pp. 245–250, 2017.View at: Publisher Site | Google Scholar

26. P. Wicks, D. L. Keininger, M. P. Massagli et al., “Perceived benefits of sharing health data between people with epilepsy on an online platform,” *Epilepsy & Behavior*, vol. 23, no. 1, pp. 16–23, 2012.View at: Publisher Site | Google Scholar

27. Social Media Likes Healthcare: From Marketing to Social Business, 2017, http://www.pwc.com/us/en/health-industries/publications/health-care-social-media.jhtml.

28. R. Batool, A. M. Khattak, J. M. Hashmi, and S. Lee, “Precise tweet classification and sentiment analysis,” in *Proceedings of the 12th International Conference on Computer and Information Science (ICIS), 2013 IEEE/ACIS*, IEEE, Niigata, Japan, June 2013.View at: Google Scholar

29. S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, “Social media analytics—challenges in topic discovery, data collection, and data preparation,” *International Journal of Information Management*, vol. 39, pp. 156–168, 2018.View at: Google Scholar

30. F. Emmert-Streib, O. P. Yli-Harja, M. Dehmer, and F. Emmert-Streib, “Data analytics applications for streaming data from social media: what to predict?” *Frontiers in Big Data*, vol. 1, p. 2, 2018.View at: Publisher Site | Google Scholar

31. O. Loyola-González, A. López-Cuevas, M. A. Medina-Pérez et al., “Fusing pattern discovery and visual analytics approaches in tweet propagation,” *Information Fusion*, vol. 46, pp. 91–101, 2018.View at: Publisher Site | Google Scholar

32. S. Petrovic, M. Osborne, V. Lavrenko et al., “RT to win! predicting message propagation in twitter,” in *Proceedings of the Fifth International Conference on Weblogs and Social Media*, vol. 13, pp. 586–589, Barcelona, Catalonia, Spain, July 2011.View at: Google Scholar

33. H. Li, D. Caragea, C. Caragea, and N. Herndon, “Disaster response aided by tweet classification with a domain adaptation approach,” *Journal of Contingencies and Crisis Management*, vol. 26, no. 1, pp. 16–27, 2018.View at: Publisher Site | Google Scholar

34. J. Chen, R. Nairn, L. Nelson et al., “Short and tweet: experiments on recommending content from information streams,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, Atlanta, GA, USA, April 2010.View at: Google Scholar

35. F. Abel, Q. Gao, G.-J. Houben, and K. Tao, “Analyzing user modeling on twitter for personalized news recommendations,” in *Proceedings of the International Conference on User Modeling, Adaptation, and Personalization*, Springer Berlin Heidelberg, Girona, Spain, July 2011.View at: Google Scholar

36. F. Abel, Q. Gao, G.-J. Houben, and K. Tao, “Semantic enrichment of Twitter posts for user profile construction on the social web,” in *Proceedings of the Extended Semantic Web Conference*, Springer Berlin Heidelberg, Heraklion, Crete, Greece, May 2011.View at: Google Scholar

37. R. Plutchick. “Emotions: A general psychoevolutionary theory.” In K.R. Scherer & P. Ekman (Eds) Approaches to emotion. Hillsdale, NJ; Lawrence Ealrbaum Associates, 1984.

38. P. Basile, V. Basile, M. Nissim, N. Novielli, V. Patti. “Sentiment Analysis of Microblogging Data”. To appear in Encyclopedia of Social Network Analysis and Mining, Springer. In press.

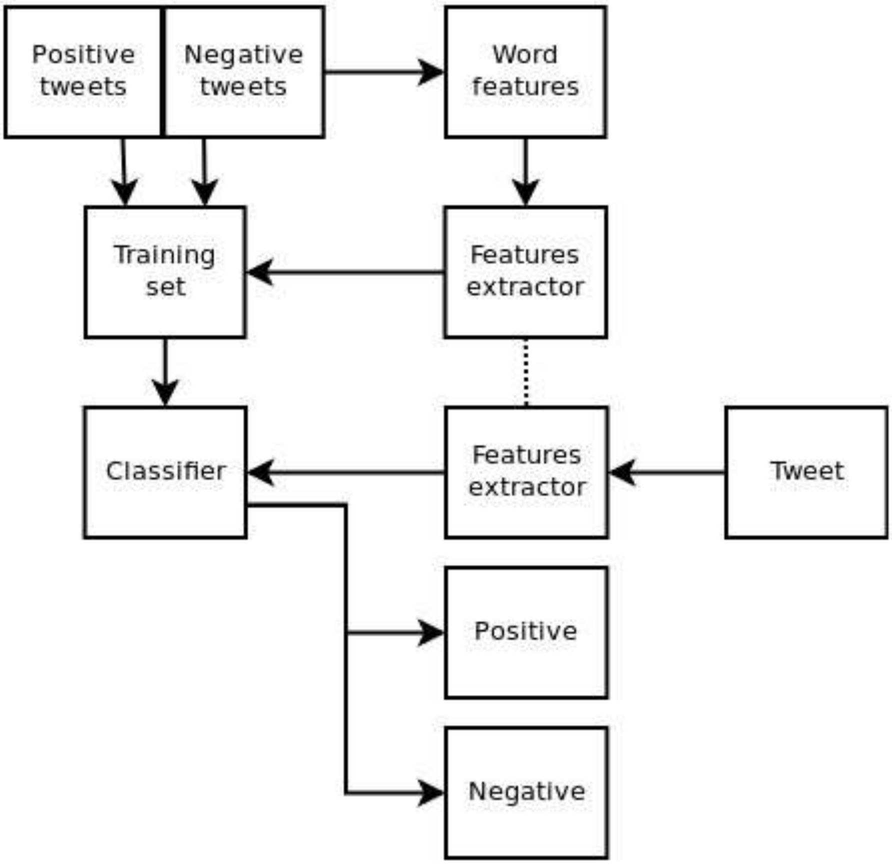
39. Johan Bollen, Huina Mao, and Alberto Pepe, “Modeling public mood and emotion: Twitter sentiment and socioeconomic phenomena.,” in International AAAI Conference on Weblogs and Social Media (ICWSM’11), 2011.

40. Cortes, Corinna; Vapnik, Vladimir N. (1995). "Supportvector networks". Machine Learning. 20 (3): 273–297.

41. Russell, Stuart; Norvig, Peter (2003) [1995]. Artificial Intelligence: A Modern Approach (2nd ed.). Prentice Hall. ISBN 978-0137903955.

42. Greene, William H. (2012). Econometric Analysis (Seventh ed.). Boston: Pearson Education. pp. 803–806. ISBN 978-0-273-75356-8.

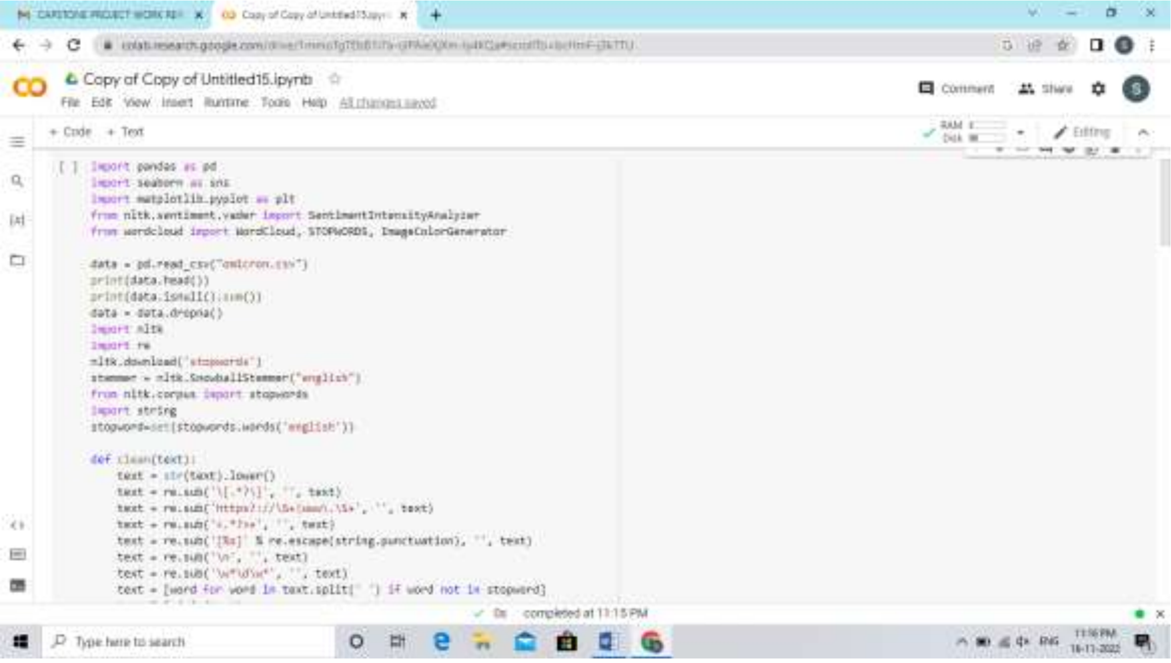
5.2 Data Flow Diagram



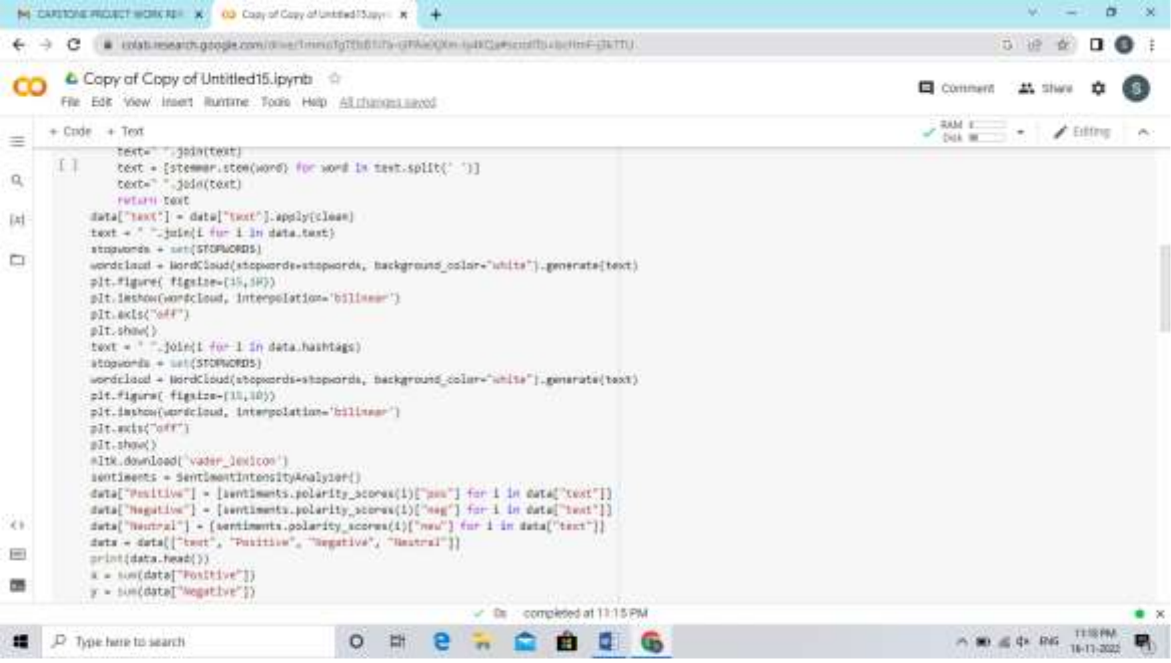
CHAPTER 6

6. IMPLEMENTATION

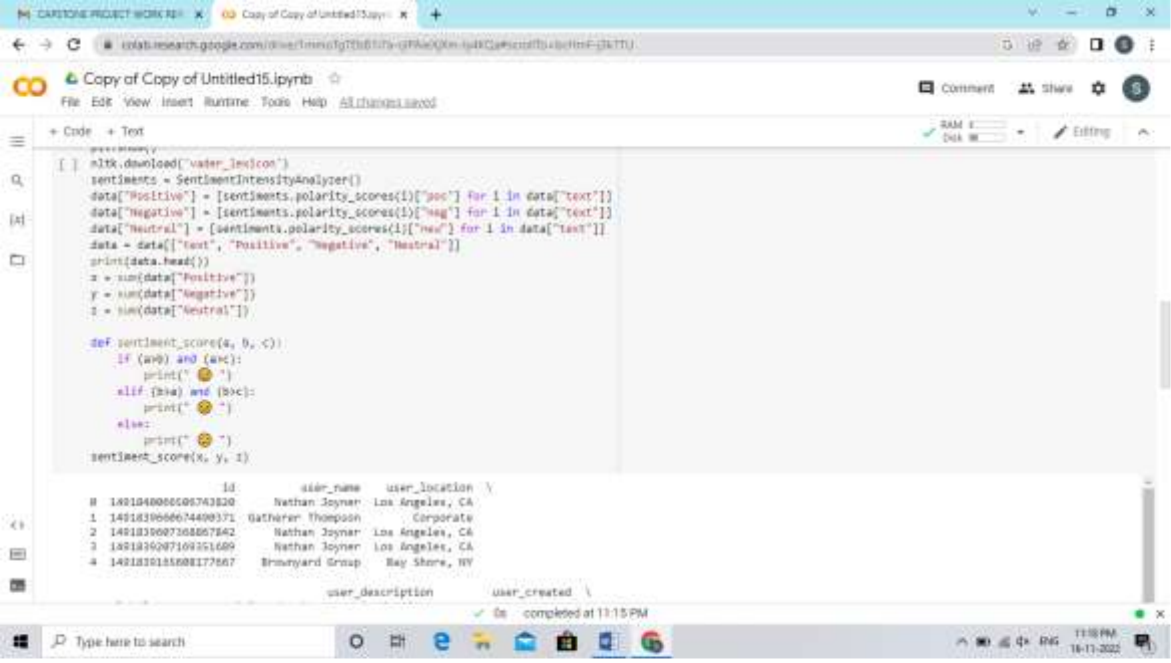
6.1 CODING


Copy of Copy of Untitled15.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
Comment Share
Editing
0s completed at 11:15 PM
Type here to search

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator

data = pd.read_csv("omicron.csv")
print(data.head())
print(data.isnull().sum())
data = data.dropna()
import nltk
import re
nltk.download('stopwords')
stemmer = nltk.SnowballStemmer("english")
from nltk.corpus import stopwords
import string
stopword=set(stopwords.words('english'))

def clean(text):
    text = str(text).lower()
    text = re.sub('\[.*?\]', '', text)
    text = re.sub('https?://\S+|www\.\S+', '', text)
    text = re.sub('<.*?>+', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
    text = re.sub('\n', '', text)
    text = re.sub('\w*\d\w*', '', text)
    text = [word for word in text.split(' ') if word not in stopword]
```


Copy of Copy of Untitled15.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
Comment Share
Editing
0s completed at 11:15 PM
Type here to search

```
    text=" ".join(text)
    text = [stemmer.stem(word) for word in text.split(' ')]
    text=" ".join(text)
    return text
data["text"] = data["text"].apply(clean)
text = " ".join(i for i in data.text)
stopwords = set(STOPWORDS)
wordcloud = WordCloud(stopwords=stopwords, background_color="white").generate(text)
plt.figure( figsize=(15,10))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
text = " ".join(i for i in data.hashtags)
stopwords = set(STOPWORDS)
wordcloud = WordCloud(stopwords=stopwords, background_color="white").generate(text)
plt.figure( figsize=(15,10))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
nltk.download('vader_lexicon')
sentiments = SentimentIntensityAnalyzer()
data["Positive"] = [sentiments.polarity_scores(i)["pos"] for i in data["text"]]
data["Negative"] = [sentiments.polarity_scores(i)["neg"] for i in data["text"]]
data["Neutral"] = [sentiments.polarity_scores(i)["neu"] for i in data["text"]]
data = data[["text", "Positive", "Negative", "Neutral"]]
print(data.head())
x = sum(data["Positive"])
y = sum(data["Negative"])
```



Copy of Copy of Untitled15.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
Comment Share
RAM Disk
Editing

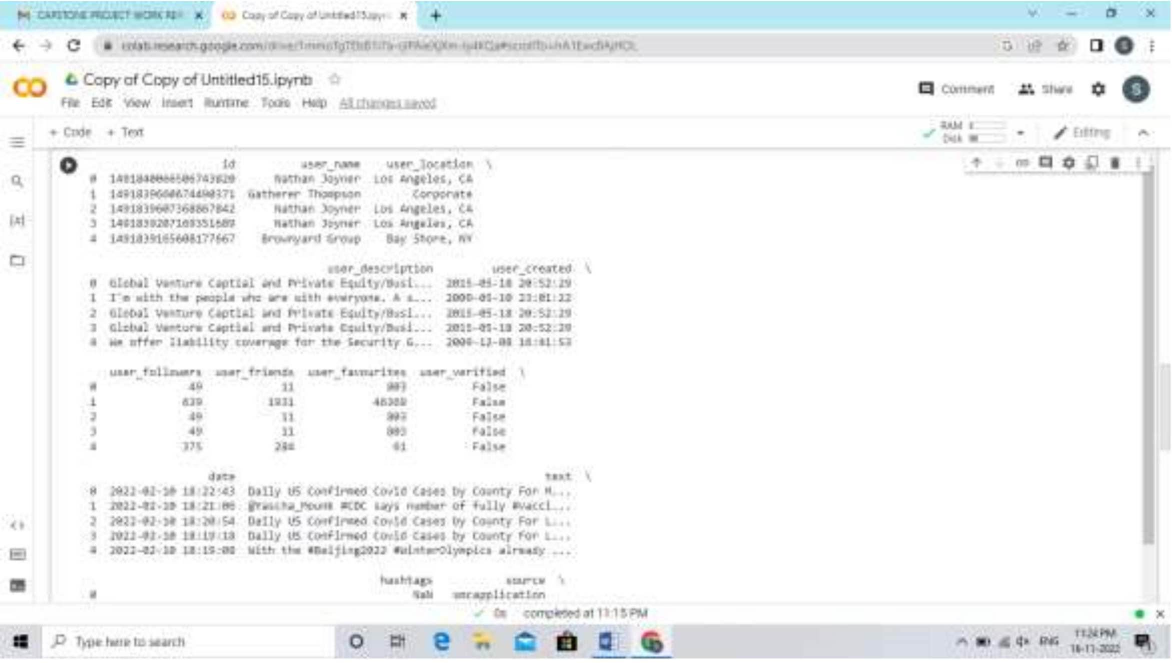
```
nltk.download("vader_lexicon")
sentiments = SentimentIntensityAnalyzer()
data["Positive"] = [sentiments.polarity_scores(i)["pos"] for i in data["text"]]
data["Negative"] = [sentiments.polarity_scores(i)["neg"] for i in data["text"]]
data["Neutral"] = [sentiments.polarity_scores(i)["neu"] for i in data["text"]]
data = data[["text", "Positive", "Negative", "Neutral"]]
print(data.head())
x = sum(data["Positive"])
y = sum(data["Negative"])
z = sum(data["Neutral"])

def sentiment_score(a, b, c):
    if (a>b) and (a>c):
        print(" 😊 ")
    elif (b>a) and (b>c):
        print(" 😟 ")
    else:
        print(" 😐 ")
sentiment_score(x, y, z)
```

```
                    id         user_name    user_location  \
0  1491840866506743820      Nathan Joyner  Los Angeles, CA
1  1491839666674490371  Gatherer Thompson        Corporate
2  1491839607368867842      Nathan Joyner  Los Angeles, CA
3  1491839207169351689      Nathan Joyner  Los Angeles, CA
4  1491839155808177667    Brownyard Group    Bay Shore, NY

                          user_description         user_created  \
```

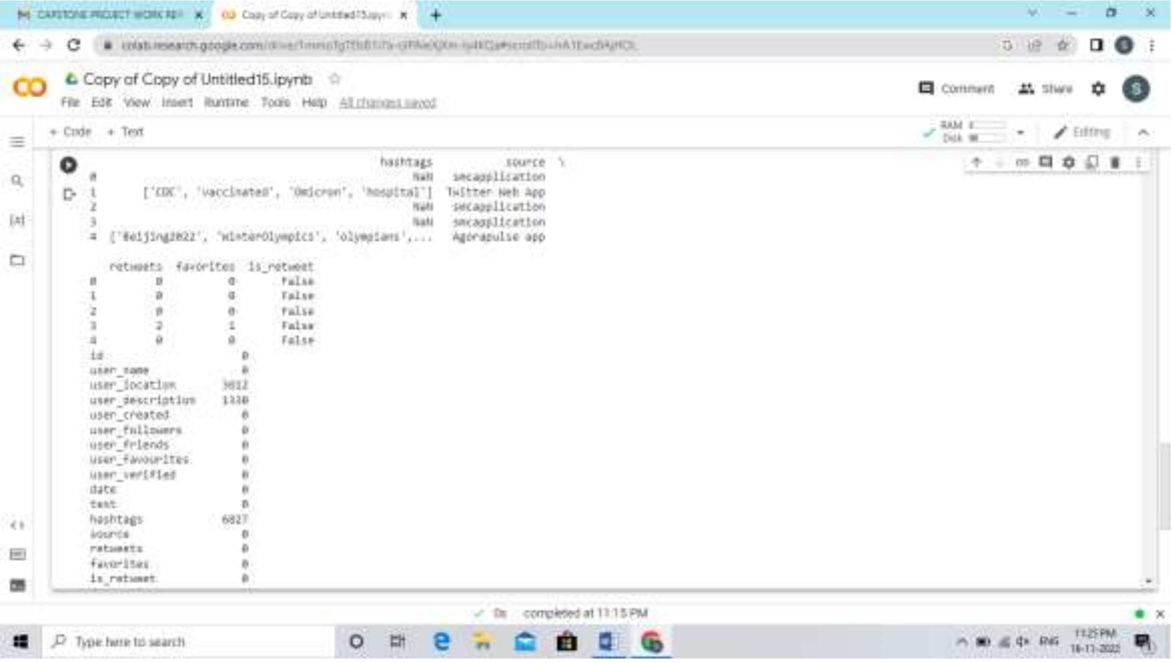
✓ 0s completed at 11:15 PM

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator

data = pd.read_csv("omicron.csv")
print(data.head())
print(data.isnull().sum())
data = data.dropna()
import nltk
import re
nltk.download('stopwords')
stemmer = nltk.SnowballStemmer("english")
from nltk.corpus import stopwords
import string
stopword=set(stopwords.words('english'))

def clean(text):
    text = str(text).lower()
    text = re.sub('\[.*?\]', '', text)
    text = re.sub('https?://\S+|www\.\S+', '', text)
    text = re.sub('<.*?>+', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
    text = re.sub('\n', '', text)
    text = re.sub('\w*\d\w*', '', text)
    text = [word for word in text.split(' ') if word not in stopword]
    text=" ".join(text)
    text = [stemmer.stem(word) for word in text.split(' ')]
    text=" ".join(text)
    return text
data["text"] = data["text"].apply(clean)
text = " ".join(i for i in data.text)
stopwords = set(STOPWORDS)
wordcloud = WordCloud(stopwords=stopwords, background_color="white").generate(text)
plt.figure( figsize=(15,10))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
text = " ".join(i for i in data.hashtags)
stopwords = set(STOPWORDS)
wordcloud = WordCloud(stopwords=stopwords, background_color="white").generate(text)
plt.figure( figsize=(15,10))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
nltk.download('vader_lexicon')
sentiments = SentimentIntensityAnalyzer()
data["Positive"] = [sentiments.polarity_scores(i)["pos"] for i in data["text"]]
data["Negative"] = [sentiments.polarity_scores(i)["neg"] for i in data["text"]]
data["Neutral"] = [sentiments.polarity_scores(i)["neu"] for i in data["text"]]
data = data[["text", "Positive", "Negative", "Neutral"]]
print(data.head())
x = sum(data["Positive"])
y = sum(data["Negative"])
z = sum(data["Neutral"])

def sentiment_score(a, b, c):
    if (a>b) and (a>c):
        print(" 😊 ")
    elif (b>a) and (b>c):
        print(" 😫 ")
    else:
        print(" 🙂 ")
sentiment_score(x, y, z)
```

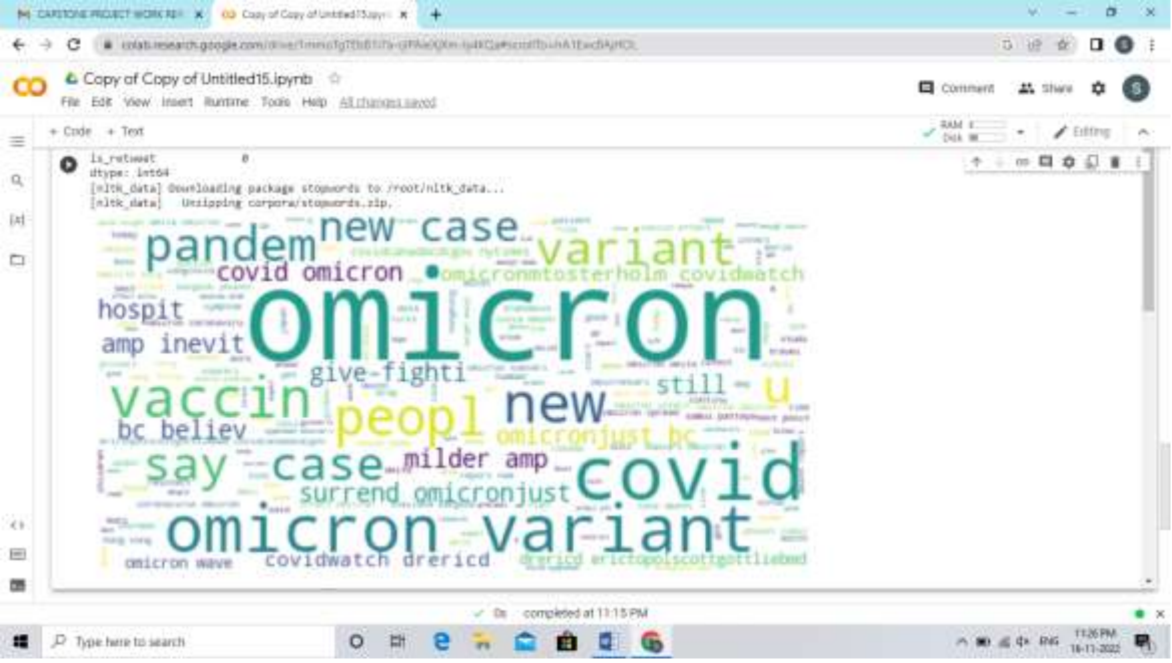
6.2 OUTPUT


Copy of Copy of Untitled15.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
Comment Share
completed at 11:15 PM
Type here to search

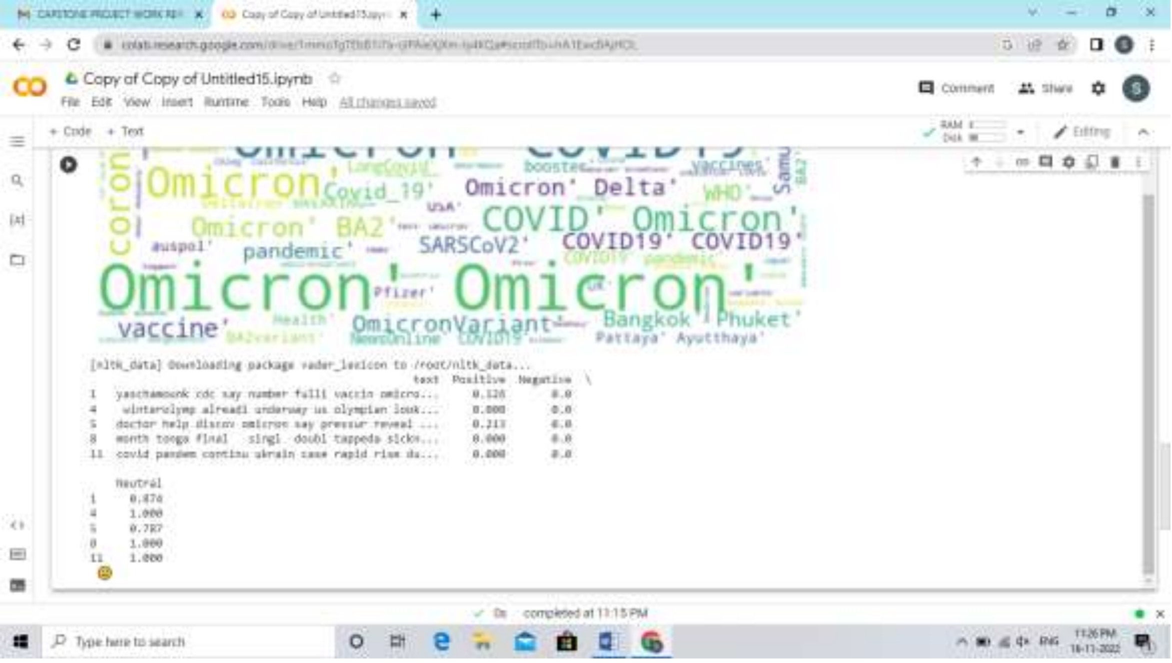
```
                    id         user_name    user_location  \
0  1491840866506743820     Nathan Joyner  Los Angeles, CA
1  1491839660674490371  Gatherer Thompson        Corporate
2  1491839607368867842     Nathan Joyner  Los Angeles, CA
3  1491839287103351689     Nathan Joyner  Los Angeles, CA
4  1491839165608177667   Brownyard Group    Bay Shore, NY

                                    user_description         user_created  \
0  Global Venture Captial and Private Equity/Busi...  2015-05-18 20:52:29
1  I'm with the people who are with everyone. A s...  2009-05-10 23:01:22
2  Global Venture Captial and Private Equity/Busi...  2015-05-18 20:52:29
3  Global Venture Captial and Private Equity/Busi...  2015-05-18 20:52:29
4  We offer liability coverage for the Security G...  2009-12-08 16:41:53

   user_followers  user_friends  user_favourites  user_verified  \
0              49            11              803          False
1             639          1931            48388          False
2              49            11              803          False
3              49            11              803          False
4             375           284               61          False

                  date                                               text  \
0  2022-02-10 18:22:43  Daily US Confirmed Covid Cases by County For M...
1  2022-02-10 18:21:06  @Vascha_Mouse #CDC says number of fully #vacci...
2  2022-02-10 18:20:54  Daily US Confirmed Covid Cases by County For L...
3  2022-02-10 18:19:18  Daily US Confirmed Covid Cases by County For L...
4  2022-02-10 18:19:08  With the #Beijing2022 #WinterOlympics already ...

                                           hashtags           source  \
0                                               NaN   smcapplication
```


Copy of Copy of Untitled15.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
Comment Share
completed at 11:15 PM
Type here to search

```
                                           hashtags           source  \
0                                               NaN   smcapplication
1       ['CDC', 'vaccinated', 'Omicron', 'hospital']  Twitter Web App
2                                               NaN   smcapplication
3                                               NaN   smcapplication
4  ['Beijing2022', 'WinterOlympics', 'olympians',...    Agorapulse app

   retweets  favorites  is_retweet
0         0          0       False
1         0          0       False
2         0          0       False
3         2          1       False
4         0          0       False
id                     0
user_name              0
user_location       3612
user_description    1330
user_created           0
user_followers         0
user_friends           0
user_favourites        0
user_verified          0
date                   0
text                   0
hashtags            6827
source                 0
retweets               0
favorites              0
is_retweet             0
```



```
is_retweet        0
dtype: int64
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```



```
[nltk_data] Downloading package vader_lexicon to /root/nltk_data...
                                                text  Positive  Negative  \
1    yaschamounk cdc say number fulli vaccin omicro...     0.126       0.0
4     winterolymp alreadi underway us olympian look...     0.000       0.0
5    doctor help discov omicron say pressur reveal ...     0.213       0.0
8    month tonga final   singl  doubl tappeda sickn...     0.000       0.0
11   covid pandem continu ukrain case rapid rise du...     0.000       0.0

    Neutral
1     0.874
4     1.000
5     0.787
8     1.000
11    1.000
```

*****