Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Engineering
Academic Year 2022-2023

Sentiment Analysis of Tweets during FIFA 2022

Web Intelligence Laboratory

By

Jazib Dawre **60004190054**
Junaid Girkar **60004190057**
Kanaad Deshpande **60004190058**

Guide(s):

Prof. Sindhu Nair

Lecturer, DJSCE

Experiment 8

Date of Performance : 20/04/23 **Date of Submission:** 27/04/23

SAP ID: 60004190057 **Name : Junaid Girkar**

Div: A **Batch : A2**

Aim of Experiment

Apply analytics to social media activity (Using FB, Twitter, Instagram or any social media dataset)

Theory / Algorithm / Conceptual Description

Social Media Analytics

Social media analytics is the ability to gather and find meaning in data gathered from social channels to support business decisions — and measure the performance of actions based on those decisions through social media.

Social media analytics is broader than metrics such as likes, follows, retweets, previews, clicks, and impressions gathered from individual channels. It also differs from reporting offered by services that support marketing campaigns such as LinkedIn or Google Analytics.

Social media analytics uses specifically designed software platforms that work similarly to web search tools. Data about keywords or topics is retrieved through search queries or web 'crawlers' that span channels. Fragments of text are returned, loaded into a database, categorized and analysed to derive meaningful insights.

Social media analytics includes the concept of social listening. Listening is monitoring social channels for problems and opportunities. Social media analytics tools typically incorporate listening into more comprehensive reporting that involves listening and performance analysis.

Social media analytics helps companies address these experiences and use them to:

1. Spot trends related to offerings and brands.
2. Understand conversations — what is being said and how it is being received.
3. Derive customer sentiment towards products and services.
4. Gauge response to social media and other communications.
5. Identify high-value features for a product or service.
6. Uncover what competitors are saying and its effectiveness.
7. Map how third-party partners and channels may affect performance.

Key capabilities of effective social media analytics

Natural language processing and machine learning technologies identify entities and relationships in unstructured data — information not pre-formatted to work with data analytics. Virtually all social media content is unstructured. These technologies are critical to deriving meaningful insights.

Segmentation is a fundamental need in social media analytics. It categorizes social media participants by geography, age, gender, marital status, parental status and other demographics. It can help identify influencers in those categories. Messages, initiatives and responses can be better tuned and targeted by understanding who is interacting on key topics.

Behavior analysis is used to understand the concerns of social media participants by assigning behavioral types such as user, recommender, prospective user and detractor. Understanding these roles helps develop targeted messages and responses to meet, change or deflect their perceptions.

Sentiment analysis measures the tone and intent of social media comments. It typically involves natural language processing technologies to help understand entities and relationships to reveal positive, negative, neutral or ambivalent attributes.

Share of voice analyzes prevalence and intensity in conversations regarding brand, products, services, reputation and more. It helps determine key issues and important topics. It also helps classify discussions as positive, negative, neutral or ambivalent.

Clustering analysis can uncover hidden conversations and unexpected insights. It makes associations between keywords or phrases that appear together frequently and derives new topics, issues and opportunities. The people that make baking soda, for example, discovered new uses and opportunities using clustering analysis.

Dashboards and visualization charts, graphs, tables and other presentation tools summarize and share social media analytics findings — a critical capability for communicating and acting on what has been learned. They also enable users to grasp meaning and insights more quickly and look deeper into specific findings without advanced technical skills.

FIFA Tweets Dataset ⚽🏆

Football is one of the most loved sports worldwide. The FIFA World Cup, a global football sporting event that takes place every four years, is in Qatar this year. This dataset contains 30,000 tweets from the first day of the FIFA World Cup 2022.

Data Collection

The dataset was created using the Snscrape and the cardiffnlp/twitter-roberta-base-sentiment-latest model in Hugging Face Hub.

Data Preprocessing

The dataset includes tweets in English containing the hashtag #WorldCup2022. For data preprocessing, we used a tokenizer for the cardiffnlp/twitter-roberta-base-sentiment-latest model and the following function:

```
def preprocess(text):
    new_text = []
    for t in text.split(" "):
        t = '@user' if t.startswith('@') and len(t) > 1 else t
        t = 'http' if t.startswith('http') else t
        new_text.append(t)
    return " ".join(new_text)
```

Data Storage

The collected tweets have been consolidated into a single dataset & shared as a Comma Separated Values file, "fifa_world_cup_2022_tweets.csv".

Content

The dataset contains as following columns:

1. Date Created
2. Number of Likes
3. Source of Tweet
4. Tweet
5. Sentiment

License

License: CC0 1.0 Universal (CC0 1.0) - Public Domain Dedication

No Copyright - You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission.

Experiment 8

April 18, 2023

[ ]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import datetime

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

```
/kaggle/input/fifa-world-cup-2022-tweets/fifa_world_cup_2022_tweets.csv
```

[ ]:
```
df = pd.read_csv("/kaggle/input/fifa-world-cup-2022-tweets/
↪fifa_world_cup_2022_tweets.csv")
```

0.1 Exploratory Data Analysis

[ ]:
```
df.head()
```

[ ]:

	Unnamed: 0	Date Created	Number of Likes	Source of Tweet	Tweet	Sentiment
0	0	2022-11-20 23:59:21+00:00	4	Twitter Web App	What are we drinking today @TucanTribe \n@MadB…	neutral
1	1	2022-11-20 23:59:01+00:00	3	Twitter for iPhone	Amazing @CanadaSoccerEN #WorldCup2022 launch …	positive
2	2	2022-11-20 23:58:41+00:00	1	Twitter for iPhone	Worth reading while watching #WorldCup2022 htt…	positive
3	3	2022-11-20 23:58:33+00:00	1	Twitter Web App	Golden Maknae shinning bright\n\nhttps://t.co/…	positive
4	4	2022-11-20 23:58:28+00:00	0	Twitter for Android	If the BBC cares so much about human rights, h…	

```
4  negative
```

[ ]:
```
df['Date Created'] =  pd.to_datetime(df['Date Created'], format='%Y-%m-%d %H:%M:
  ↪%S')

df.info( )
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22524 entries, 0 to 22523
Data columns (total 6 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Unnamed: 0       22524 non-null  int64
 1   Date Created     22524 non-null  datetime64[ns, UTC]
 2   Number of Likes  22524 non-null  int64
 3   Source of Tweet  22524 non-null  object
 4   Tweet            22524 non-null  object
 5   Sentiment        22524 non-null  object
dtypes: datetime64[ns, UTC](1), int64(2), object(3)
memory usage: 1.0+ MB
```

[ ]:
```
df.shape
```

[ ]:
```
(22524, 6)
```

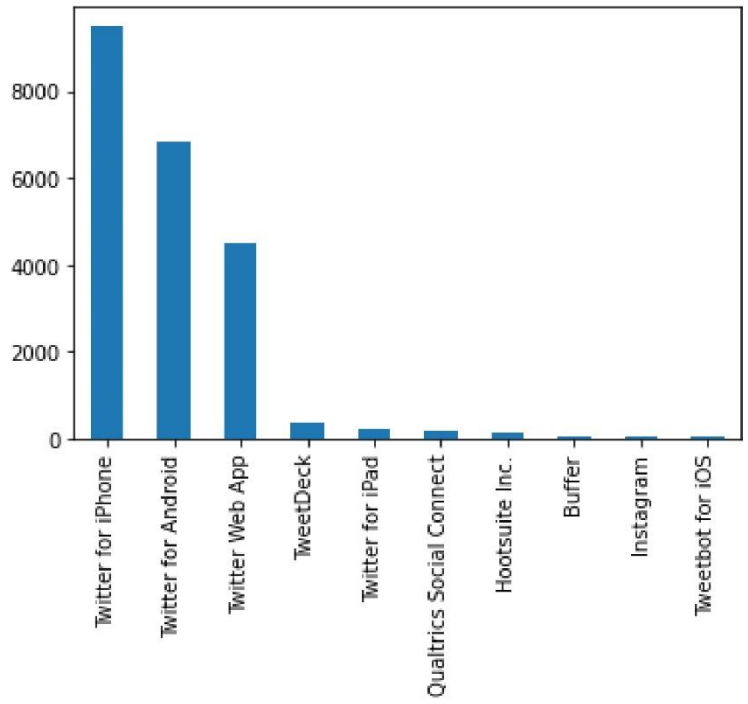
[ ]:
```
df.isnull( ).sum( )
```

[ ]:
```
Unnamed: 0         0
Date Created       0
Number of Likes    0
Source of Tweet    0
Tweet              0
Sentiment          0
dtype: int64
```

[ ]:
```
df['Source of Tweet'].value_counts().head(10)
```

[ ]:
```
Twitter for iPhone          9507
Twitter for Android         6820
Twitter Web App             4505
TweetDeck                    386
Twitter for iPad             240
Qualtrics Social Connect     165
Hootsuite Inc.               146
Buffer                        71
Instagram                     62
Tweetbot for iOS              52
Name: Source of Tweet, dtype: int64
```

[ ]:
```
source_t =  df['Source of Tweet'].value_counts()
source_t.head(10).plot.bar()

plt.show()
```



So turns out iphone users tweet more about the world cup, followed by android

[ ]:
```
most_liked = df.sort_values(by=['Number of Likes'], ascending=False)
most_liked.head(10)
```

[ ]:

	Unnamed: 0	Date Created	Number of Likes	\
1287	1287	2022-11-20 19:39:11+00:00	316867	
17359	17359	2022-11-20 12:41:31+00:00	31517	
21267	21267	2022-11-20 06:17:00+00:00	20016	
528	528	2022-11-20 22:20:55+00:00	5847	
20073	20073	2022-11-20 09:05:12+00:00	5555	
15332	15332	2022-11-20 14:57:02+00:00	5484	
20101	20101	2022-11-20 09:02:09+00:00	5430	

3

```
212           212 2022-11-20 23:06:08+00:00             3172
1635         1635 2022-11-20 19:00:00+00:00             2428
5761         5761 2022-11-20 16:23:30+00:00             2370

            Source of Tweet  \
1287     Twitter for iPhone
17359  Twitter Media Studio
21267       Twitter Web App
528               TweetDeck
20073    Twitter for iPhone
15332       Twitter Web App
20101             TweetDeck
212    Twitter Media Studio
1635    Sprinklr Publishing
5761    Twitter for Android

                                                  Tweet Sentiment
1287   I can’t express my gratitude and happiness for…  positive
17359  Football Legend Eric Cantona reminds football …   neutral
21267   #BitKeep FootBall Carnival Main Event is co…  positive
528    WATCH: BTS's #Jungkook Performs At #WorldCup20…   neutral
20073  The   Leo with Louis Vuitton   #WorldCup2022 h…  positive
15332  Need them at the World Cup opening ceremony  \…   neutral
20101  WATCH: #BTS's #Jungkook Shares His Support For…   neutral
212    "It's Palestine…"\n\nLebanese fans refuse to…  negative
1635   .@reymysterio, @fightbobby, @BiancaBelairWWE a…   neutral
5761   Qatar's Goalkeeper  \n\n#WorldCup2022 https://…   neutral
```

0.1.1 Most Liked Tweets

[ ]:
```
df.iloc[1287].Tweet
```

[ ]:
```
'I can’t express my gratitude and happiness for my participation in the biggest
event of all times the World Cup in my country #Qatar with the talented Jung
kook , thank you for everything\n#WorldCup \n#WorldCup2022 \n#JUNGKOOKxFIFA
\n#JungkookAtFIFAWorldCup \n@BTS_twt \n@bts_bighit https://t.co/PebpCj9uN2'
```

[ ]:
```
df.iloc[20101].Tweet
```

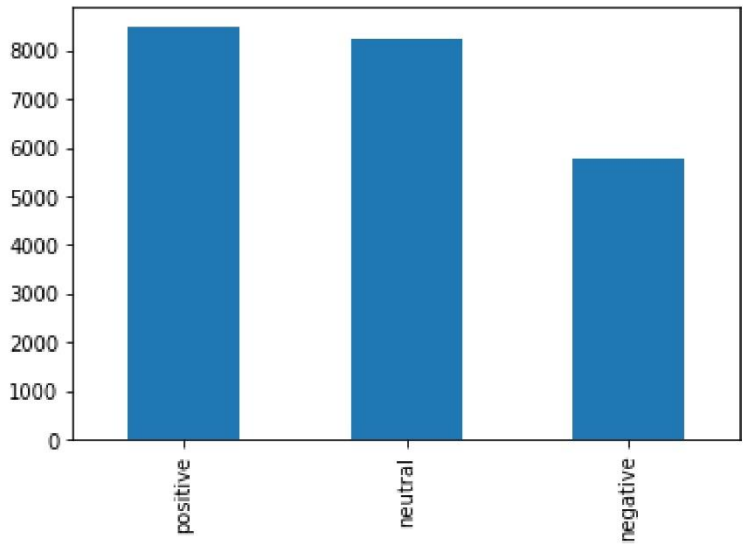
[ ]:
```
"WATCH: #BTS's #Jungkook Shares His Support For South Korea's National Soccer
Team At The #WorldCup2022 \nhttps://t.co/SHS3bmi60F https://t.co/SqSOKFKxXe"
```

[ ]:
```
df.iloc[212].Tweet
```

[ ]:
```
'"It\'s Palestine…"\n\nLebanese fans refuse to be interviewed by an Israeli TV
reporter after he said that he is from \'Israel\' at #WorldCup2022 in Qatar.
https://t.co/shdWDn4ZvC'
```

0.2 Sentiment Analysis

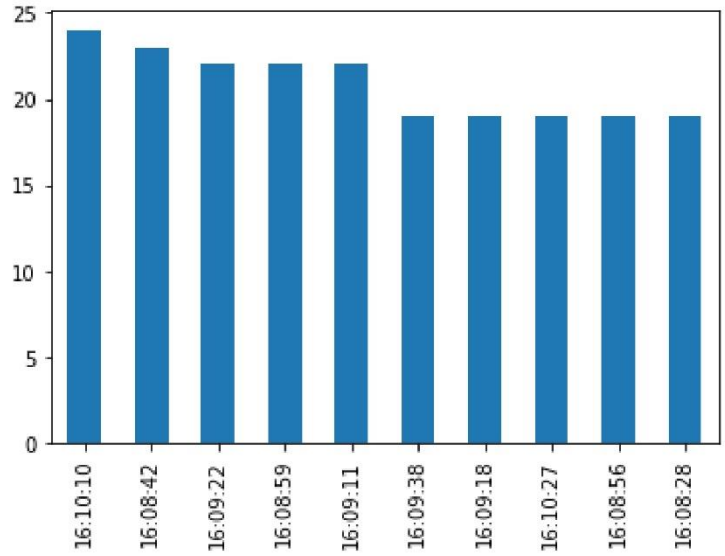
```
[ ]: sentiment =  df['Sentiment'].value_counts()
     sentiment.head().plot.bar()

     plt.show()
```




```
[ ]: df['Date Created'] = pd.to_datetime(df['Date Created']).dt.time
     df['Date Created']
```

```
[ ]: 0        23:59:21
     1        23:59:01
     2        23:58:41
     3        23:58:33
     4        23:58:28
              ...
     22519    00:00:21
     22520    00:00:03
     22521    00:00:01
     22522    00:00:00
     22523    00:00:00
     Name: Date Created, Length: 22524, dtype: object
```

[ ]:
```
ppd =  df['Date Created'].value_counts()
ppd.head(10).plot.bar()

plt.show()
```



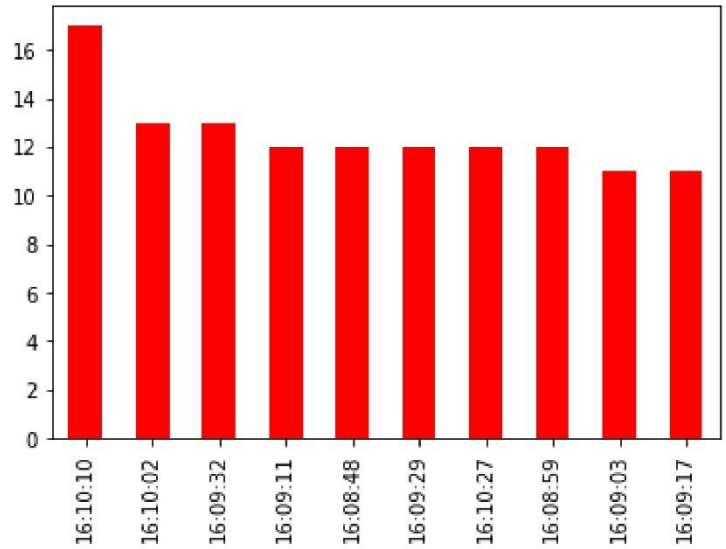
[ ]:
```
from wordcloud import WordCloud, STOPWORDS

comment_words = ''
stopwords = set(STOPWORDS)
stopwords.add('https')
stopwords.add('t')
stopwords.add('co')

# iterate through the csv file
for val in df['Tweet']:

    # typecaste each val to string
    val = str(val)

    # split the value
    tokens = val.split()

    # Converts each token into lowercase
```

```
    for i in range(len(tokens)):
        tokens[i] = tokens[i].lower()


    comment_words += " ".join(tokens)+" "

wordcloud = WordCloud(width = 800, height = 800,
                background_color ='white',
                stopwords = stopwords,
                min_font_size = 10).generate(comment_words)

# plot the WordCloud image
plt.figure(figsize = (8, 8), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad = 0)

plt.show()
```


```
[ ]: neg_tweets = df[df['Sentiment'] == 'negative']
     neg_tweets = neg_tweets['Date Created'].value_counts()
     neg_tweets.head(10).plot.bar(color=['red'])

     neg_tweets.head()
```

```
[ ]: 16:10:10    17
     16:10:02    13
     16:09:32    13
     16:09:11    12
     16:08:48    12
     Name: Date Created, dtype: int64
```

8



[ ]:
```
negdf = pd.DataFrame()
negdf['negative'] = neg_tweets

postweets = df[df['Sentiment'] == 'positive']
posdf = pd.DataFrame()

posdf['positive'] = postweets['Date Created'].value_counts()
f_df = pd.DataFrame()

f_df = pd.merge(negdf, posdf, left_index=True, right_index=True)
```

[ ]:
```
f_df = f_df.sort_index(ascending=True)
f_df
```

[ ]:

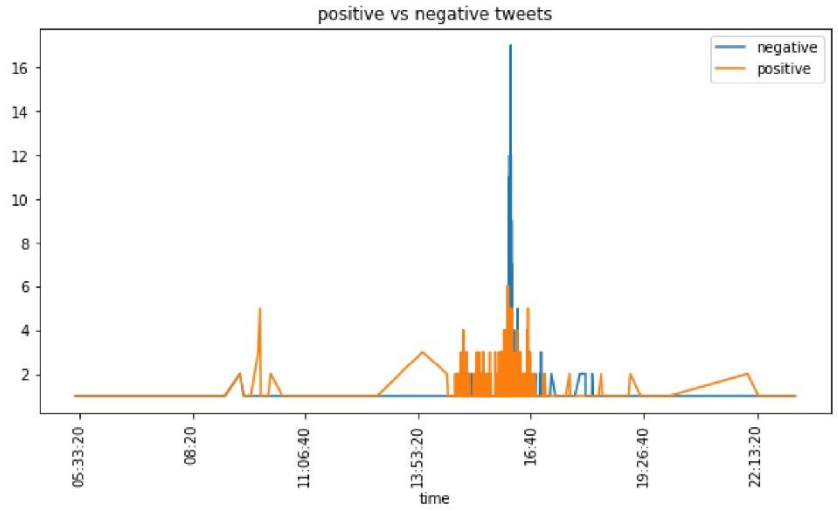
	negative	positive
05:27:09	1	1
05:57:24	1	1
06:55:09	1	1
07:20:04	1	1
07:37:27	1	1
…	…	…
20:03:55	1	1
22:00:01	1	2

9

```
22:15:52        1         1
22:49:59        1         1
23:09:59        1         1

[1053 rows x 2 columns]
```

```
[ ]: f_df.plot(figsize=(10, 5), rot=90, title="positive vs negative tweets")
```

```
[ ]: <AxesSubplot:title={'center':'positive vs negative tweets'}, xlabel='time'>
```



Conclusion

Sentiment analysis is a powerful tool that allows generalization of user feedback for various demographics with multiple levels of granularity. The most common implementation over social media data allows companies and events to evaluate user feedback and consumer sentiment towards products and policies without conducting explicit surveys. It also removes bias that is inherent with organizational surveying by directly reading user interactions in a social web of communication.