ankit-data-science-project

October 14, 2024

0.1 Steps to be followed:

Importing necessary libraries Loading datasets Data preprocessing Exploratory Data Analysis Sentiment Analysis Conclusion

0.2 Dataset: US Election 2024 Tweets and hashtag

0.2.1 Dataset features:

created_at: Date and time of tweet posted tweet_id: Tweet's unique ID tweet: Full tweet text likes: Number of likes retweet_count: Number of retweets source: Utility used to post the tweet user_id: User ID of tweet creator user_name: Username of tweet creator user_screen_name: Screen name of tweet creator user_description: Self-description by tweet creator user_join_date: Join date of tweet creator user_followers_count: Followers count on tweet creator user_location: Address was given on tweeter's profile lat: Latitude parsed from user_location long: Longitude parsed from user_location city: City parsed from user_location country: Country parsed from user_location state: State parsed from user_location state_code: State code parsed from user_location collected_at: Date and time tweet data was mined from Twitter Let's begin with the implementation.

0.2.2 Importing datasets

[1]:
```
# Import Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import plotly.express as px

# Libraries for Sentiment Analysis
import re
import nltk
from nltk.corpus import stopwords
from nltk.corpus import wordnet
from nltk.stem import WordNetLemmatizer
from textblob import TextBlob
from wordcloud import WordCloud

# to avoid warnings
import warnings
warnings.filterwarnings('ignore')
```

[2]:
```
# reading datasets
trump = pd.read_csv("hashtag_donaldtrump.csv", lineterminator='\n')
print(trump.head(3))
```

```
            created_at      tweet_id  \
0  2020-10-15 00:00:01  1.316529e+18
1  2020-10-15 00:00:01  1.316529e+18
2  2020-10-15 00:00:02  1.316529e+18

                                               tweet  likes  retweet_count  \
0  #Elecciones2020 | En #Florida: #JoeBiden dice …    0.0            0.0
1  Usa 2020, Trump contro Facebook e Twitter: cop…   26.0            9.0
2  #Trump: As a student I used to hear for years,…    2.0            1.0

             source      user_id           user_name user_screen_name  \
0         TweetDeck  360666534.0  El Sol Latino News  elsollatinonews
1  Social Mediaset   331617619.0             Tgcom24  MediasetTgcom24
2   Twitter Web App    8436472.0              snarke           snarke

                                  user_description  …  \
0   Noticias de interés para latinos de la costa…  …
1  Profilo ufficiale di Tgcom24: tutte le notizie…  …
2  Will mock for food! Freelance writer, blogger,…  …

  user_followers_count                user_location        lat        long  \
0               1860.0  Philadelphia, PA / Miami, FL  25.774270  -80.193660
1            1067661.0                           NaN        NaN         NaN
2               1185.0                      Portland  45.520247 -122.674195

       city                   country      continent    state state_code  \
0       NaN  United States of America  North America  Florida         FL
1       NaN                       NaN            NaN      NaN        NaN
2  Portland  United States of America  North America   Oregon         OR

                   collected_at
0           2020-10-21 00:00:00
1  2020-10-21 00:00:00.373216530
2  2020-10-21 00:00:00.746433060

[3 rows x 21 columns]
```

0.2.3 Let's have a look at all the features in this dataset

[3]:
```
# Display all the columns in the DataFrame
print(trump.columns)
```

```
Index(['created_at', 'tweet_id', 'tweet', 'likes', 'retweet_count', 'source',
       'user_id', 'user_name', 'user_screen_name', 'user_description',
```

```
       'user_join_date', 'user_followers_count', 'user_location', 'lat',
       'long', 'city', 'country', 'continent', 'state', 'state_code',
       'collected_at'],
      dtype='object')
```

[4]:
```
biden = pd.read_csv("hashtag_joebiden.csv", lineterminator='\n')
print(biden.head(2))
```

```
            created_at      tweet_id  \
0  2020-10-15 00:00:01  1.316529e+18
1  2020-10-15 00:00:18  1.316529e+18

                                               tweet  likes  retweet_count  \
0  #Elecciones2020 | En #Florida: #JoeBiden dice …    0.0            0.0
1  #HunterBiden #HunterBidenEmails #JoeBiden #Joe…    0.0            0.0

             source      user_id           user_name user_screen_name  \
0         TweetDeck  360666534.0  El Sol Latino News  elsollatinonews
1  Twitter for iPad  809904438.0         Cheri A.       Biloximeemaw

                                  user_description  …  \
0   Noticias de interés para latinos de la costa…  …
1  Locked and loaded Meemaw. Love God, my family …  …

  user_followers_count                user_location       lat      long  \
0               1860.0  Philadelphia, PA / Miami, FL  25.77427 -80.19366
1               6628.0                           NaN       NaN       NaN

   city                    country      continent    state state_code  \
0   NaN  United States of America  North America  Florida         FL
1   NaN                        NaN            NaN      NaN        NaN

                   collected_at
0           2020-10-21 00:00:00
1  2020-10-21 00:00:00.517827283

[2 rows x 21 columns]
```

0.2.4 Data Assessment:

for example, studying the shape of data and what it tells, checking variables and their data types

[5]:
```
print(trump.shape)
print(biden.shape)
```

```
(970919, 21)
(776886, 21)
```

[6]:
```
# Getting trump dataset information
trump.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 970919 entries, 0 to 970918
Data columns (total 21 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   created_at            970919 non-null  object
 1   tweet_id              970919 non-null  float64
 2   tweet                 970919 non-null  object
 3   likes                 970919 non-null  float64
 4   retweet_count         970919 non-null  float64
 5   source                970043 non-null  object
 6   user_id               970919 non-null  float64
 7   user_name             970897 non-null  object
 8   user_screen_name      970919 non-null  object
 9   user_description      869651 non-null  object
 10  user_join_date        970919 non-null  object
 11  user_followers_count  970919 non-null  float64
 12  user_location         675957 non-null  object
 13  lat                   445719 non-null  float64
 14  long                  445719 non-null  float64
 15  city                  227187 non-null  object
 16  country               442748 non-null  object
 17  continent             442765 non-null  object
 18  state                 320620 non-null  object
 19  state_code            300425 non-null  object
 20  collected_at          970919 non-null  object
dtypes: float64(7), object(14)
memory usage: 155.6+ MB
```

[7]:
```
# Getting biden dataset information
biden.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 776886 entries, 0 to 776885
Data columns (total 21 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   created_at            776886 non-null  object
 1   tweet_id              776886 non-null  float64
 2   tweet                 776886 non-null  object
 3   likes                 776886 non-null  float64
 4   retweet_count         776886 non-null  float64
 5   source                776173 non-null  object
 6   user_id               776886 non-null  float64
 7   user_name             776861 non-null  object
```

```
 8   user_screen_name      776886 non-null  object
 9   user_description      694877 non-null  object
 10  user_join_date        776886 non-null  object
 11  user_followers_count  776886 non-null  float64
 12  user_location         543092 non-null  object
 13  lat                   355293 non-null  float64
 14  long                  355293 non-null  float64
 15  city                  186872 non-null  object
 16  country               353779 non-null  object
 17  continent             353797 non-null  object
 18  state                 260195 non-null  object
 19  state_code            244609 non-null  object
 20  collected_at          776886 non-null  object
dtypes: float64(7), object(14)
memory usage: 124.5+ MB
```

0.2.5 Data Preprocessing

[8]:
```
# creating a new column 'candidate' todifferentiate
# between tweets of Trump and Biden upon concatination
trump['candidate'] = 'trump'

# biden dataframe
biden['candidate'] = 'biden'

# combining the dataframes
data = pd.concat([trump, biden])

# FInal data shape
print('Final Data Shape :', data.shape)

# View the first 2 rows
print("\nFirst 2 rows:")
print(data.head(3))
```

```
Final Data Shape : (1747805, 22)

First 2 rows:
            created_at      tweet_id  \
0  2020-10-15 00:00:01  1.316529e+18
1  2020-10-15 00:00:01  1.316529e+18
2  2020-10-15 00:00:02  1.316529e+18

                                               tweet  likes  retweet_count  \
0  #Elecciones2020 | En #Florida: #JoeBiden dice …    0.0            0.0
1  Usa 2020, Trump contro Facebook e Twitter: cop…   26.0            9.0
2  #Trump: As a student I used to hear for years,…    2.0            1.0
```

```
             source      user_id           user_name user_screen_name  \
0         TweetDeck  360666534.0  El Sol Latino News  elsollatinonews
1  Social Mediaset   331617619.0             Tgcom24  MediasetTgcom24
2   Twitter Web App    8436472.0              snarke           snarke

                                  user_description  …  \
0   Noticias de interés para latinos de la costa…  …
1  Profilo ufficiale di Tgcom24: tutte le notizie…  …
2  Will mock for food! Freelance writer, blogger,…  …

                  user_location        lat        long      city  \
0  Philadelphia, PA / Miami, FL  25.774270  -80.193660       NaN
1                           NaN        NaN         NaN       NaN
2                      Portland  45.520247 -122.674195  Portland

                    country      continent    state state_code  \
0  United States of America  North America  Florida         FL
1                       NaN            NaN      NaN        NaN
2  United States of America  North America   Oregon         OR

                    collected_at candidate
0            2020-10-21 00:00:00     trump
1  2020-10-21 00:00:00.373216530     trump
2  2020-10-21 00:00:00.746433060     trump

[3 rows x 22 columns]
```

0.2.6 Data Cleaning:

Dropping missing values

[9]:
```
# dropping null values if they exist
data.dropna(inplace=True)
```

[10]:
```
data['country'].value_counts()
```

[10]:
```
country
United States of America    182382
United Kingdom               31869
India                        20931
France                       19996
Germany                      18534
Canada                       16250
The Netherlands               8491
Australia                     8330
Spain                         5254
Brazil                        4211
Pakistan                      3704
```

```
Italy                          2966
Ireland                        2587
Bangladesh                     2036
Mexico                         1972
Belgium                        1962
Nigeria                        1848
South Africa                   1648
United Arab Emirates           1521
Switzerland                    1494
Peru                           1031
Lebanon                        1002
Argentina                       872
Ecuador                         824
Colombia                        565
Honduras                        508
Venezuela                       431
New Zealand                     384
Poland                          340
Uruguay                         237
Lithuania                       198
Bolivia                         194
El Salvador                     171
Oman                            105
Philippines                      74
Trinidad and Tobago              65
Papua New Guinea                 60
Kuwait                           22
Sudan                            21
Burkina Faso                     20
Syria                            19
Suriname                         19
Slovakia                         19
Guatemala                        16
Côte d'Ivoire                    16
Laos                              8
Libya                             5
South Sudan                       5
Guyana                            4
Somalia                           2
Cameroon                          1
Name: count, dtype: int64
```

[11]:
```
data['country'] = data['country'].replace({'United States of America': "US",
                                           'United␣
 ↪States': "US"})
```

0.2.7 Exploratory Data Analysis (EDA)

[12]:
```
# Group the data by 'candidate' and count the
# number of tweets for each candidate
tweets_count = data.groupby('candidate')['tweet'].count().reset_index()

# Interactive bar chart
fig = px.bar(tweets_count, x='candidate', y='tweet', color='candidate',
                        color_discrete_map={'Trump': 'pink', 'Biden': 'blue'},
                        labels={'candidate': 'Candidates', 'tweet': 'Number of␣
 ↪Tweets'},
                        title='Tweets for Candidates')

# Show the chart
fig.show()
```

[13]:
```
# Interactive bar chart
likes_comparison = data.groupby('candidate')['likes'].sum().reset_index()
fig = px.bar(likes_comparison, x='candidate', y='likes', color='candidate',
                        color_discrete_map={'Trump': 'blue', 'Biden': 'green'},
                        labels={'candidate': 'Candidate', 'likes': 'Total␣
 ↪Likes'},
                        title='Comparison of Likes')

# Update the layout with a black theme
fig.update_layout(plot_bgcolor='black',
                                paper_bgcolor='black', font_color='white')

# Show the chart
fig.show()
```

[14]:
```
# Top10 Countrywise tweets Counts
top10countries = data.groupby('country')['tweet'].count(
).sort_values(ascending=False).reset_index().head(10)
# top10countries

# Interactive bar chart
fig = px.bar(top10countries, x='country', y='tweet',
                        template='plotly_dark',
                        color_discrete_sequence=px.colors.qualitative.Dark24_r,
                        title='Top10 Countrywise tweets Counts')

# To view the graph
fig.show()
```

Tweet Counts for Each Candidate in the Top 10 Countries

Now, let us find out the number of tweets done for each candidate by all the countries.

[15]:
```
# the number of tweets done for each
# candidate by all the countries.
tweet_df = data.groupby(['country', 'candidate'])[
        'tweet'].count().reset_index()

# Candidate for top 10 country tweet
tweeters = tweet_df[tweet_df['country'].isin(top10countries.country)]

# Plot for tweet counts for each candidate
# in the top 10 countries
fig = px.bar(tweeters, x='country', y='tweet', color='candidate',
                        labels={'country': 'Country', 'tweet': 'Number of␣
 ↪Tweets',
                                              'candidate': 'Candidate'},
                        title='Tweet Counts for Each Candidate in the Top 10␣
 ↪Countries',
                        template='plotly_dark',
                        barmode='group')

# Show the chart
fig.show()
```

[16]:
```
def clean(text):
        # Remove URLs
        text = re.sub(r'https?://\S+|www\.\S+', '', str(text))

        # Convert text to lowercase
        text = text.lower()

        # Replace anything other than alphabets a-z with a space
        text = re.sub('[^a-z]', ' ', text)

        # Split the text into single words
        text = text.split()

        # Initialize WordNetLemmatizer
        lm = WordNetLemmatizer()

        # Lemmatize words and remove stopwords
        text = [lm.lemmatize(word) for word in text if word not in set(
                stopwords.words('english'))]

        # Join the words back into a sentence
        text = ' '.join(word for word in text)

        return text
```

[17]:
```
def getpolarity(text):
        return TextBlob(text).sentiment.polarity

def getsubjectivity(text):
        return TextBlob(text).sentiment.subjectivity

def getAnalysis(score):
        if score < 0:
                return 'negative'
        elif score == 0:
                return 'neutral'
        else:
                return 'positive'
```

[18]:
```
trump_tweets = data[data['candidate'] == 'trump']

# taking only U.S. country data
trump_tweets = trump_tweets.loc[trump_tweets.country == 'US']
trump_tweets = trump_tweets[['tweet']]
print(trump_tweets.head())
```

```
                                                tweet
2   #Trump: As a student I used to hear for years,…
4   You get a tie! And you get a tie! #Trump ‘s ra…
11  In 2020, #NYPost is being #censorship #CENSORE…
12  #Trump #PresidentTrump #Trump2020LandslideVict…
22  #Trump: Nobody likes to tell you this, but som…
```

[19]:
```
trump_tweets['cleantext'] = trump_tweets['tweet'].apply(clean)
print(trump_tweets.head())
```

```
                                                tweet  \
2   #Trump: As a student I used to hear for years,…
4   You get a tie! And you get a tie! #Trump ‘s ra…
11  In 2020, #NYPost is being #censorship #CENSORE…
12  #Trump #PresidentTrump #Trump2020LandslideVict…
22  #Trump: Nobody likes to tell you this, but som…

                                           cleantext
2   trump student used hear year ten year heard ch…
4                    get tie get tie trump rally iowa
11  nypost censorship censored twitter manipulate …
12  trump presidenttrump trump landslidevictory tr…
22  trump nobody like tell farmer better way worki…
```

[32]:
```
trump_tweets['subjectivity'] = trump_tweets['cleantext'].apply(getsubjectivity)
```

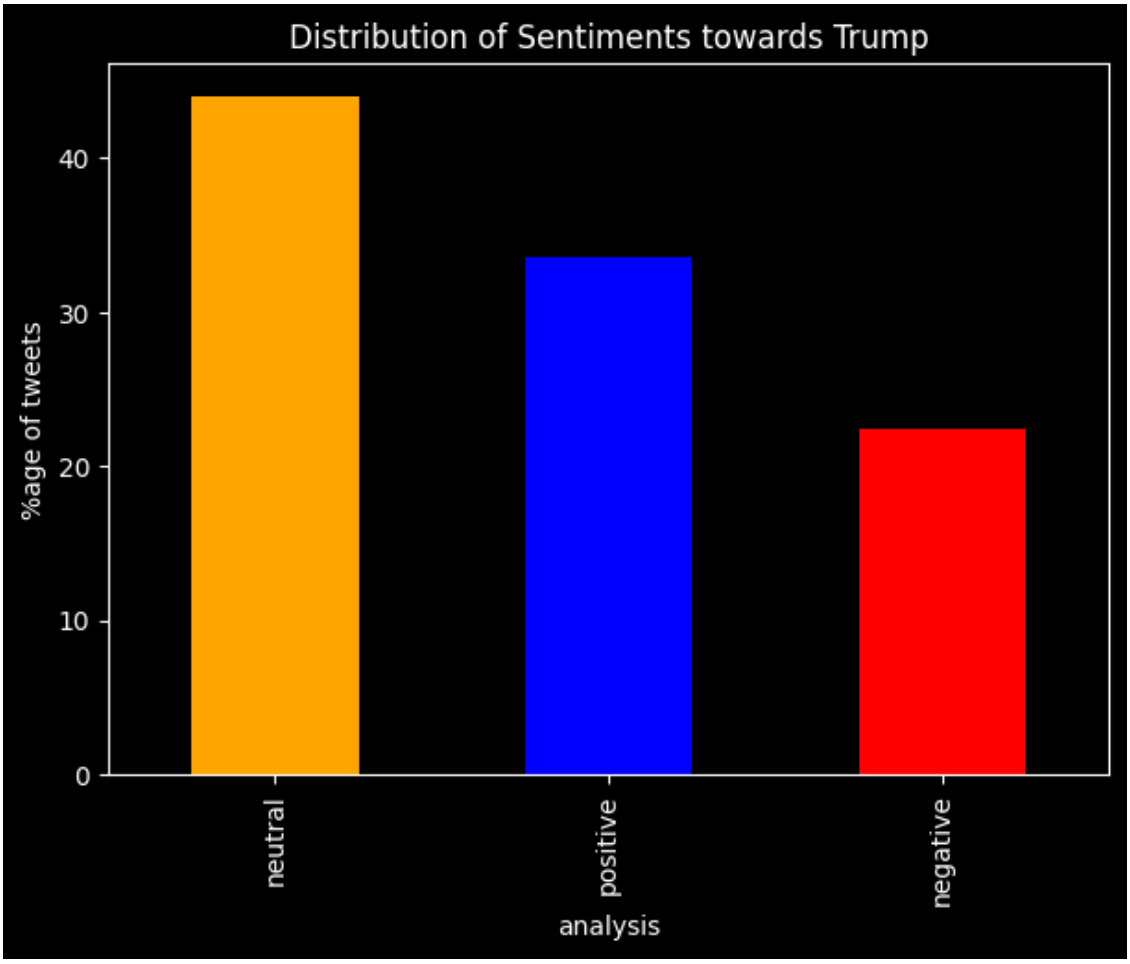
[33]:
```
trump_tweets['polarity'] = trump_tweets['cleantext'].apply(getpolarity)
```

[22]:
```
trump_tweets['analysis'] = trump_tweets['polarity'].apply(getAnalysis)
trump_tweets.head()
```

[22]:

	tweet	cleantext	subjectivity	polarity	analysis
2	#Trump: As a student I used to hear for years,…	trump student used hear year ten year heard ch…	0.333333	0.333333	positive
4	You get a tie! And you get a tie! #Trump ‘s ra…	get tie get tie trump rally iowa	0.000000	0.000000	neutral
11	In 2020, #NYPost is being #censorship #CENSORE…	nypost censorship censored twitter manipulate …	0.678571	-0.148810	negative
12	#Trump #PresidentTrump #Trump2020LandslideVict…	trump presidenttrump trump landslidevictory tr…	0.750000	0.500000	positive
22	#Trump: Nobody likes to tell you this, but som…	trump nobody like tell farmer better way worki…	0.595238	0.261905	positive

[23]:
```
# how much data is positive/negetive/neutral
plt.style.use('dark_background') # Adding black theme

# Define colors for each bar
colors = ['orange', 'blue', 'red']

plt.figure(figsize=(7, 5))
(trump_tweets.analysis.value_counts(normalize=True) * 100).plot.
 ↪bar(color=colors)
plt.ylabel("%age of tweets")
plt.title("Distribution of Sentiments towards Trump")
plt.show()
```

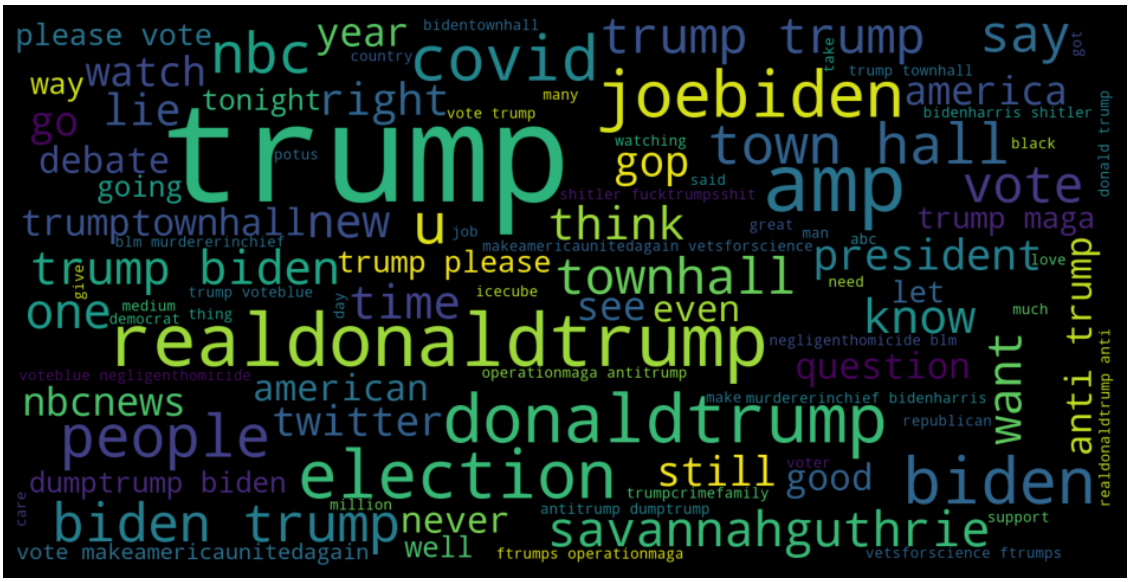


[24]:
```
!pip install wordcloud
import matplotlib.pyplot as plt
from wordcloud import WordCloud, STOPWORDS

def word_cloud(wd_list):
        stopwords = set(STOPWORDS)
        all_words = ' '.join(wd_list)
        wordcloud = WordCloud(background_color='black',
                                                stopwords=stopwords,
                                                width=1600, height=800,␣
 ↪max_words=100, max_font_size=200,
                                                colormap="viridis").
 ↪generate(all_words)
        plt.figure(figsize=(12, 10))
        plt.axis('off')
        plt.imshow(wordcloud)
```

```
word_cloud(trump_tweets['cleantext'][:5000])
```

```
[notice] A new release of pip is available: 24.0 -> 24.2
[notice] To update, run: python.exe -m pip install --upgrade pip
```

```
Requirement already satisfied: wordcloud in
c:\users\user\appdata\local\programs\python\python312\lib\site-packages (1.9.3)
Requirement already satisfied: numpy>=1.6.1 in
c:\users\user\appdata\local\programs\python\python312\lib\site-packages (from
wordcloud) (2.0.0)
Requirement already satisfied: pillow in
c:\users\user\appdata\local\programs\python\python312\lib\site-packages (from
wordcloud) (10.4.0)
Requirement already satisfied: matplotlib in
c:\users\user\appdata\local\programs\python\python312\lib\site-packages (from
wordcloud) (3.9.1)
Requirement already satisfied: contourpy>=1.0.1 in
c:\users\user\appdata\local\programs\python\python312\lib\site-packages (from
matplotlib->wordcloud) (1.2.1)
Requirement already satisfied: cycler>=0.10 in
c:\users\user\appdata\local\programs\python\python312\lib\site-packages (from
matplotlib->wordcloud) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in
c:\users\user\appdata\local\programs\python\python312\lib\site-packages (from
matplotlib->wordcloud) (4.53.1)
Requirement already satisfied: kiwisolver>=1.3.1 in
c:\users\user\appdata\local\programs\python\python312\lib\site-packages (from
matplotlib->wordcloud) (1.4.5)
Requirement already satisfied: packaging>=20.0 in
c:\users\user\appdata\local\programs\python\python312\lib\site-packages (from
matplotlib->wordcloud) (24.1)
Requirement already satisfied: pyparsing>=2.3.1 in
c:\users\user\appdata\local\programs\python\python312\lib\site-packages (from
matplotlib->wordcloud) (3.1.2)
Requirement already satisfied: python-dateutil>=2.7 in
c:\users\user\appdata\local\programs\python\python312\lib\site-packages (from
matplotlib->wordcloud) (2.9.0.post0)
Requirement already satisfied: six>=1.5 in
c:\users\user\appdata\local\programs\python\python312\lib\site-packages (from
python-dateutil>=2.7->matplotlib->wordcloud) (1.16.0)
```



[25]:
```
biden_tweets = data[data['candidate'] == 'biden']
biden_tweets = biden_tweets.loc[biden_tweets.country == 'US']
biden_tweets = biden_tweets[['tweet']]
biden_tweets
```

[25]:
```
                                                    tweet
6       In 2020, #NYPost is being #censorship #CENSORE…
17      Comments on this? "Do Democrats Understand how…
25      @RealJamesWoods #BidenCrimeFamily #JoeBiden #H…
29      Come on @ABC PLEASE DO THE RIGHT THING. Move t…
34      #realDonaldTrump addresses #JoeBiden and #Hunt…
…                                                     …
776836  #Biden     | Images   @ Santa Maria, CA.  | …
776845  Will #criticalRaceTheory become ubiquitous in …
776847   You moving near #Biden   https://t.co/1F6i1YIJ2P
776861  #election #2020Elections #trump #biden https:/…
776865  @FLOTUS I'm excited to have a FLOTUS whose vag…

[85760 rows x 1 columns]
```

[26]:
```
biden_tweets['cleantext']=biden_tweets['tweet'].apply(clean)
biden_tweets.head()
```

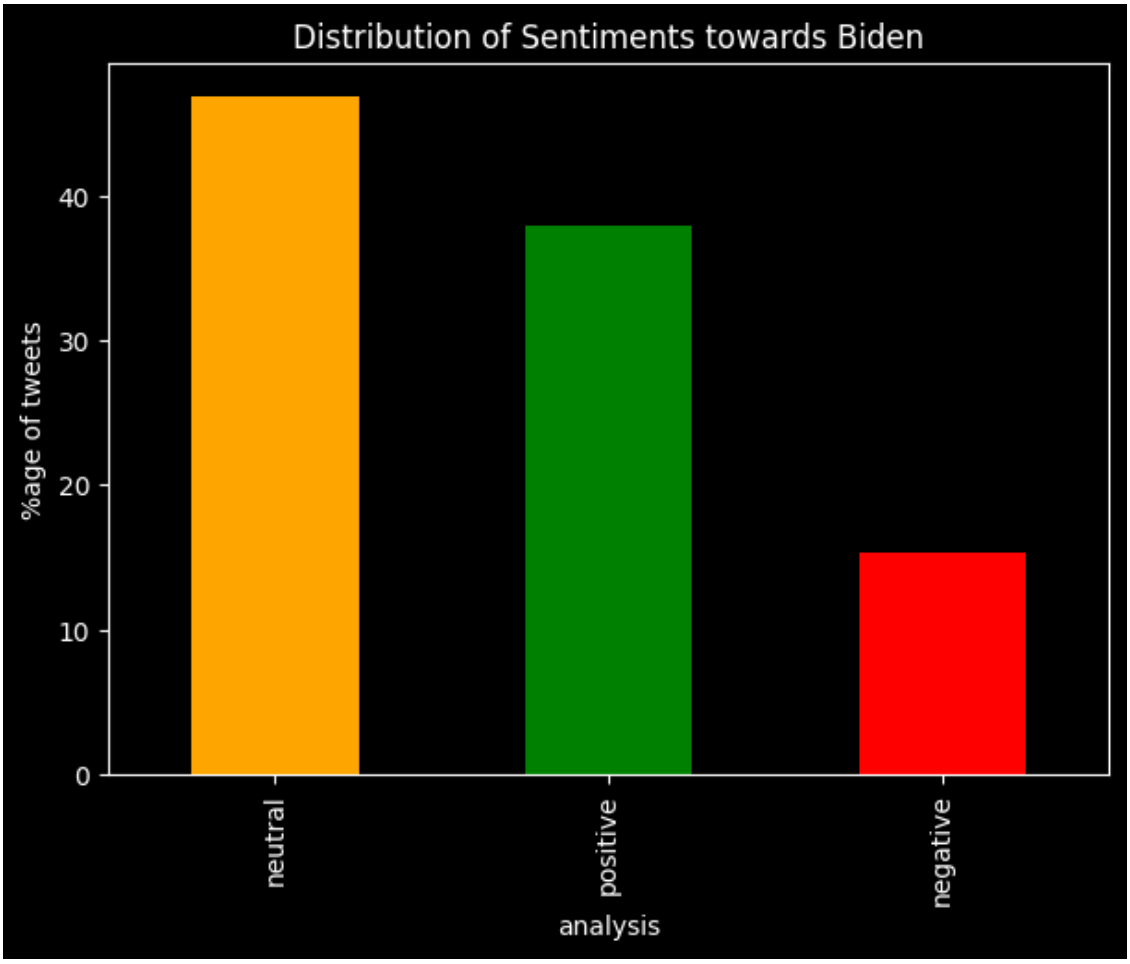
[26]:
```
                                                    tweet  \
6   In 2020, #NYPost is being #censorship #CENSORE…
17  Comments on this? "Do Democrats Understand how…
25  @RealJamesWoods #BidenCrimeFamily #JoeBiden #H…
29  Come on @ABC PLEASE DO THE RIGHT THING. Move t…
```

```
34  #realDonaldTrump addresses #JoeBiden and #Hunt…

                                                cleantext
6   nypost censorship censored twitter manipulate …
17  comment democrat understand ruthless china chi…
25  realjameswoods bidencrimefamily joebiden hunte…
29  come abc please right thing move biden town ha…
34  realdonaldtrump address joebiden hunterbiden c…
```

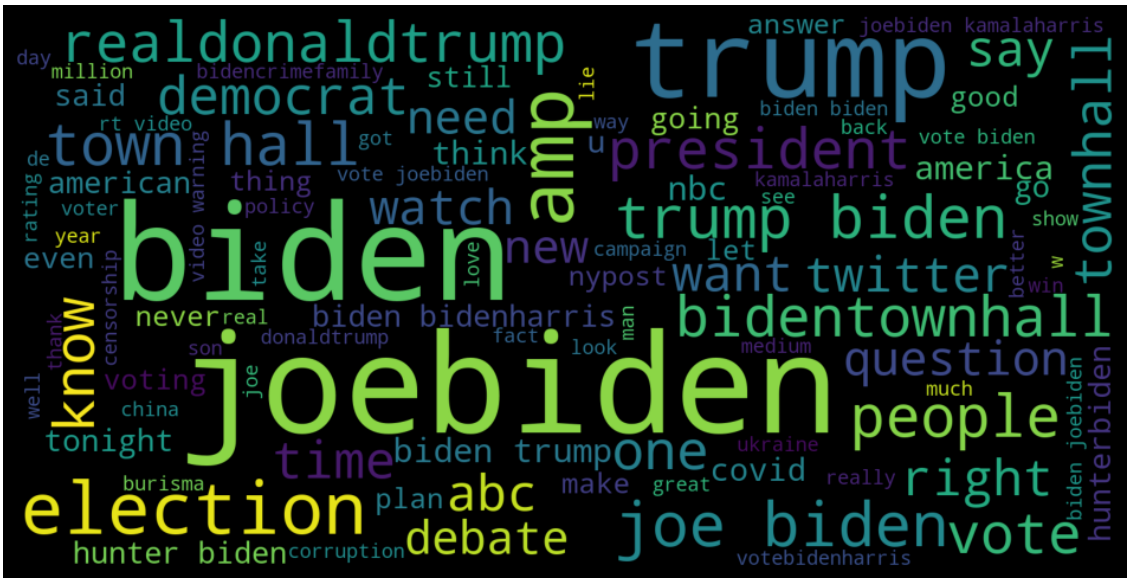
[27]:
```
biden_tweets['subjectivity'] = biden_tweets['cleantext'].apply(getsubjectivity)
biden_tweets['polarity'] = biden_tweets['cleantext'].apply(getpolarity)
biden_tweets['analysis'] = biden_tweets['polarity'].apply(getAnalysis)
biden_tweets.head()
```

[27]:
```
                                                 tweet  \
6   In 2020, #NYPost is being #censorship #CENSORE…
17  Comments on this? "Do Democrats Understand how…
25  @RealJamesWoods #BidenCrimeFamily #JoeBiden #H…
29  Come on @ABC PLEASE DO THE RIGHT THING. Move t…
34  #realDonaldTrump addresses #JoeBiden and #Hunt…

                                            cleantext  subjectivity  polarity  \
6   nypost censorship censored twitter manipulate …      0.678571 -0.148810
17  comment democrat understand ruthless china chi…      1.000000 -1.000000
25  realjameswoods bidencrimefamily joebiden hunte…      0.000000  0.000000
29  come abc please right thing move biden town ha…      0.178571  0.078571
34  realdonaldtrump address joebiden hunterbiden c…      0.000000  0.000000

    analysis
6   negative
17  negative
25   neutral
29  positive
34   neutral
```

[28]:
```
# how much data is positive/negetive/neutral
plt.style.use('dark_background')

# Define colors for each bar
colors = ['orange', 'green', 'red']

plt.figure(figsize=(7, 5))
(biden_tweets.analysis.value_counts(normalize=True) * 100).plot.
 ↪bar(color=colors)
plt.ylabel("%age of tweets")
plt.title("Distribution of Sentiments towards Biden")
plt.show()
```



[29]:
```
word_cloud(biden_tweets['cleantext'][:5000])
```



[30]:
```
trump_tweets.analysis.value_counts(normalize=True)*100
```

[30]:
```
analysis
neutral     43.995156
positive    33.566890
negative    22.437954
Name: proportion, dtype: float64
```

[31]:
```
biden_tweets.analysis.value_counts(normalize=True)*100
```

[31]:
```
analysis
neutral     46.831856
positive    37.880131
negative    15.288013
Name: proportion, dtype: float64
```

[ ]: