sentiment-analysis-checkpoint

May 24, 2024

1 Prashant Priyadarshi

1.0.1 Sentiment Analysis using NLP

[2]:

```python
# Importing Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from string import punctuation
from nltk.tokenize import word_tokenize
from nltk.stem import LancasterStemmer
from string import punctuation
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import LancasterStemmer
from nltk.stem.wordnet import WordNetLemmatizer
import re
import warnings
warnings.filterwarnings('ignore')
```

[3]:

```python
train = pd.read_csv('train_tweet.csv')
test = pd.read_csv('test_tweets.csv')

print(train.shape)
print(test.shape)
```

```
(31962, 3)
(17197, 2)
```

[4]:

```python
train.head()
```

[4]:

```
   id  label                                              tweet
0   1      0   @user when a father is dysfunctional and is s…
1   2      0  @user @user thanks for #lyft credit i can't us…
2   3      0                                bihday your majesty
3   4      0  #model   i love u take with u all the time in …
4   5      0             factsguide: society now    #motivation
```

[5]:

```
test.head()
```

[5]:

```
      id                                              tweet
0  31963  #studiolife #aislife #requires #passion #dedic…
1  31964   @user #white #supremacists want everyone to s…
2  31965  safe ways to heal your #acne!!    #altwaystohe…
3  31966  is the hp and the cursed child book up for res…
4  31967    3rd #bihday to my amazing, hilarious #nephew…
```

[6]:

```
train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 31962 entries, 0 to 31961
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   id      31962 non-null  int64
 1   label   31962 non-null  int64
 2   tweet   31962 non-null  object
dtypes: int64(2), object(1)
memory usage: 749.2+ KB
```

[7]:

```
train.isnull().any()
test.isnull().any()
```

[7]:

```
id       False
tweet    False
dtype: bool
```

[8]:

```
# checking out the negative comments from the train set

train[train['label'] == 0].head(10)
```

[8]:

```
   id  label                                              tweet
0   1      0   @user when a father is dysfunctional and is s…
1   2      0  @user @user thanks for #lyft credit i can't us…
2   3      0                              bihday your majesty
3   4      0  #model   i love u take with u all the time in …
4   5      0             factsguide: society now    #motivation
5   6      0  [2/2] huge fan fare and big talking before the…
6   7      0   @user camping tomorrow @user @user @user @use…
7   8      0  the next school year is the year for exams.ð …
8   9      0  we won!!! love the land!!! #allin #cavs #champ…
9  10      0   @user @user welcome here !  i'm   it's so #gr…
```

[9]:

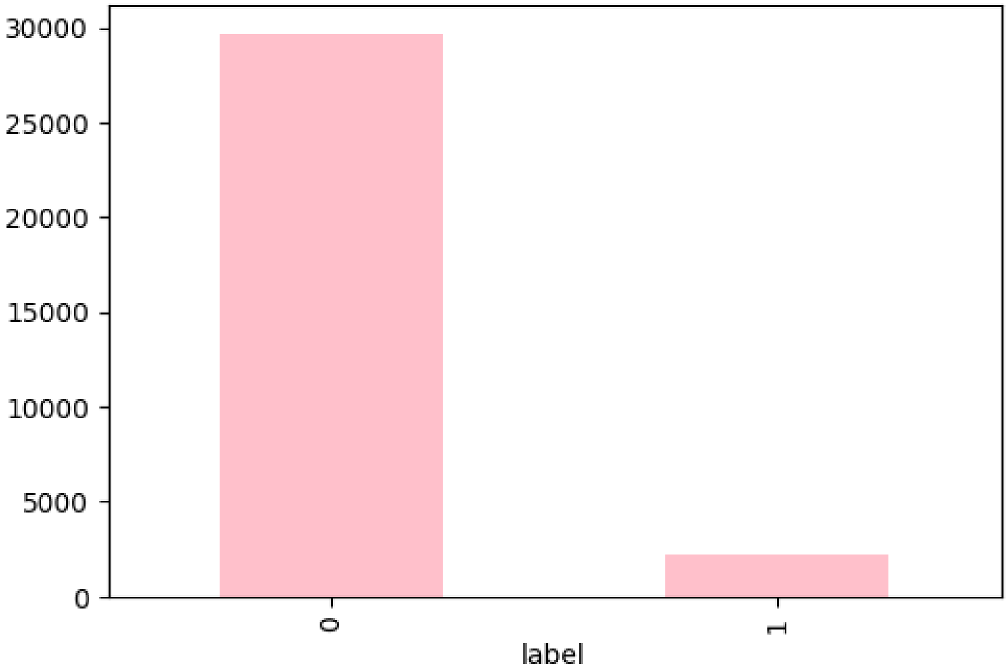
```
# checking out the postive comments from the train set
```

```
train[train['label'] == 1].head(10)
```

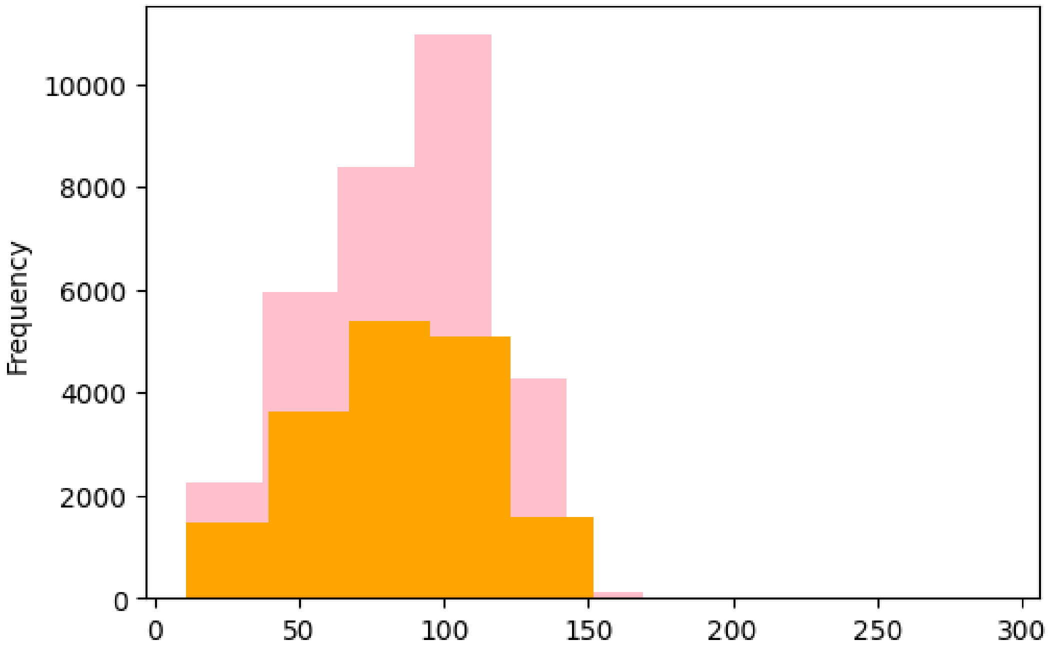
```
[9]:       id  label                                              tweet
     13    14      1  @user #cnn calls #michigan middle school 'buil…
     14    15      1  no comment!  in #australia   #opkillingbay #se…
     17    18      1                             retweet if you agree!
     23    24      1    @user @user lumpy says i am a . prove it lumpy.
     34    35      1  it's unbelievable that in the 21st century we'…
     56    57      1            @user lets fight against  #love #peace
     68    69      1  ð ©the white establishment can't have blk fol…
     77    78      1  @user hey, white people: you can call people '…
     82    83      1  how the #altright uses  &amp; insecurity to lu…
     111  112      1  @user i'm not interested in a #linguistics tha…
```

```
[10]: train['label'].value_counts().plot.bar(color = 'pink', figsize = (6, 4))
```

```
[10]: <Axes: xlabel='label'>
```



```
[11]: # checking the distribution of tweets in the data

      length_train = train['tweet'].str.len().plot.hist(color = 'pink', figsize = (6,␣
       ↪4))
      length_test = test['tweet'].str.len().plot.hist(color = 'orange', figsize = (6,␣
       ↪4))
```



[12]:
```
# adding a column to represent the length of the tweet

train['len'] = train['tweet'].str.len()
test['len'] = test['tweet'].str.len()

train.head(10)
```

[12]:
```
   id  label                                              tweet  len
0   1      0   @user when a father is dysfunctional and is s…  102
1   2      0  @user @user thanks for #lyft credit i can't us…  122
2   3      0                                bihday your majesty   21
3   4      0  #model   i love u take with u all the time in …   86
4   5      0             factsguide: society now    #motivation   39
5   6      0  [2/2] huge fan fare and big talking before the…  116
6   7      0   @user camping tomorrow @user @user @user @use…   74
7   8      0  the next school year is the year for exams.ð …  143
8   9      0  we won!!! love the land!!! #allin #cavs #champ…   87
9  10      0   @user @user welcome here !  i'm   it's so #gr…   50
```

[13]:
```
train.groupby('label').describe()
```

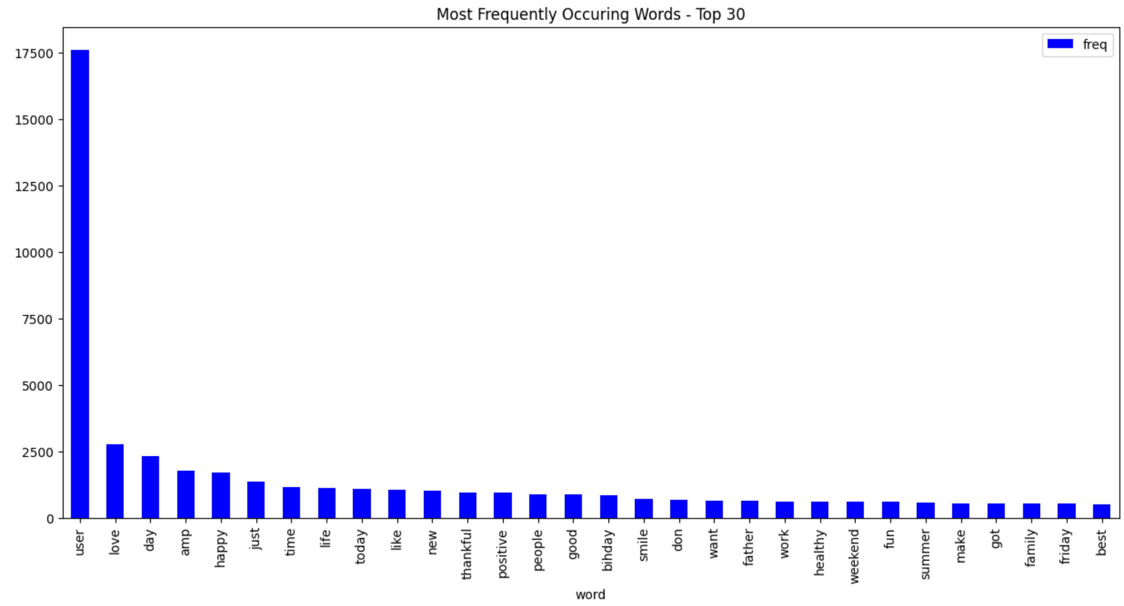
[13]:
```
            id
         count          mean          std  min      25%      50%       75%
label
0      29720.0  15974.454441  9223.783469  1.0  7981.75  15971.5  23965.25   \
1        2242.0  16074.896075  9267.955758  14.0  8075.25  16095.0  24022.00

                     len
           max     count        mean         std   min   25%   50%    75%    max
label
0      31962.0   29720.0  84.328634  29.566484  11.0  62.0  88.0  107.0  274.0
1      31961.0    2242.0  90.187779  27.375502  12.0  69.0  96.0  111.0  152.0
```

[14]:

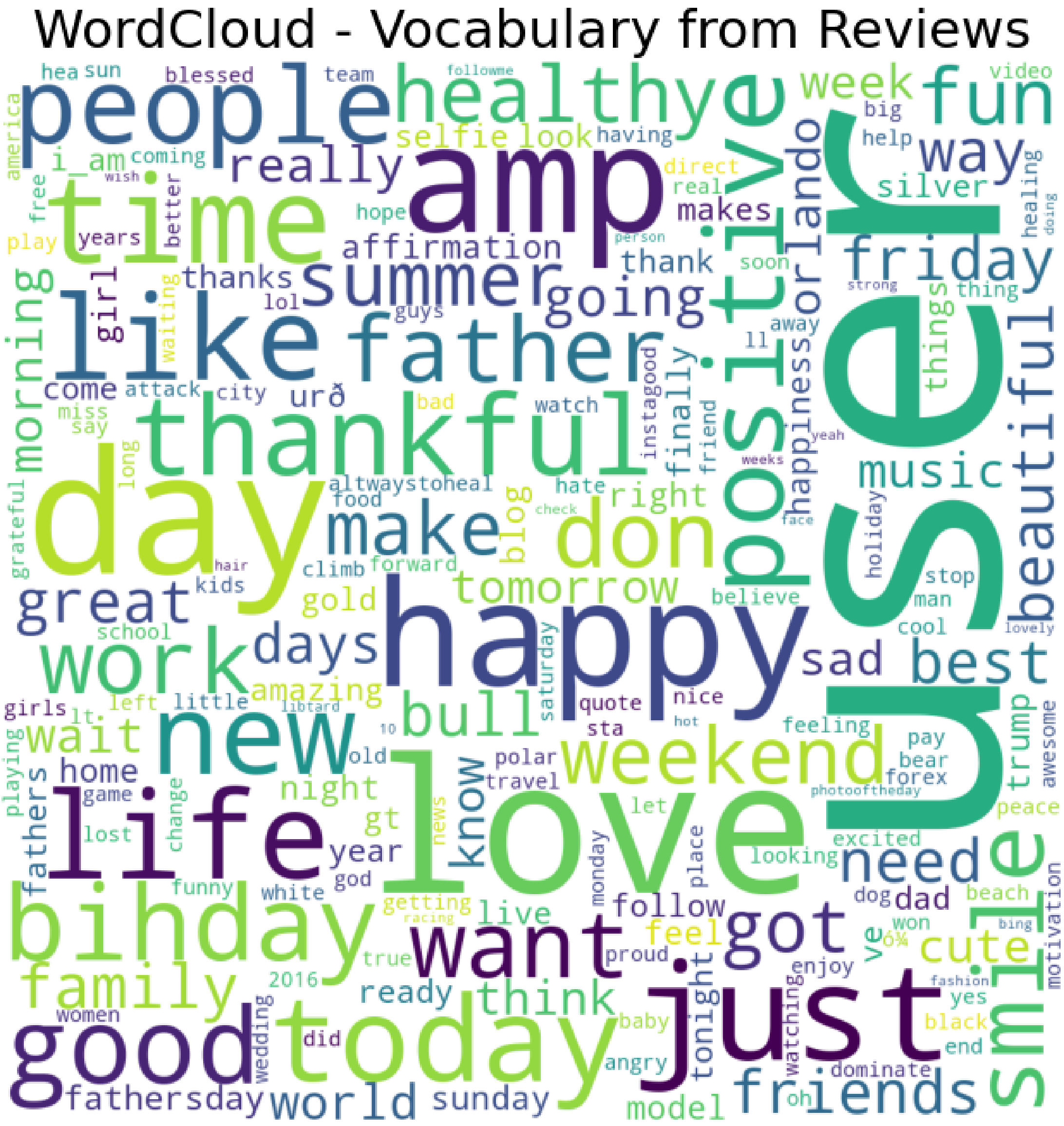
```
from sklearn.feature_extraction.text import CountVectorizer


cv = CountVectorizer(stop_words = 'english')
words = cv.fit_transform(train.tweet)

sum_words = words.sum(axis=0)

words_freq = [(word, sum_words[0, i]) for word, i in cv.vocabulary_.items()]
words_freq = sorted(words_freq, key = lambda x: x[1], reverse = True)

frequency = pd.DataFrame(words_freq, columns=['word', 'freq'])

frequency.head(30).plot(x='word', y='freq', kind='bar', figsize=(15, 7), color␣
 ↪= 'blue')
plt.title("Most Frequently Occuring Words - Top 30")
```

[14]:

```
Text(0.5, 1.0, 'Most Frequently Occuring Words - Top 30')
```

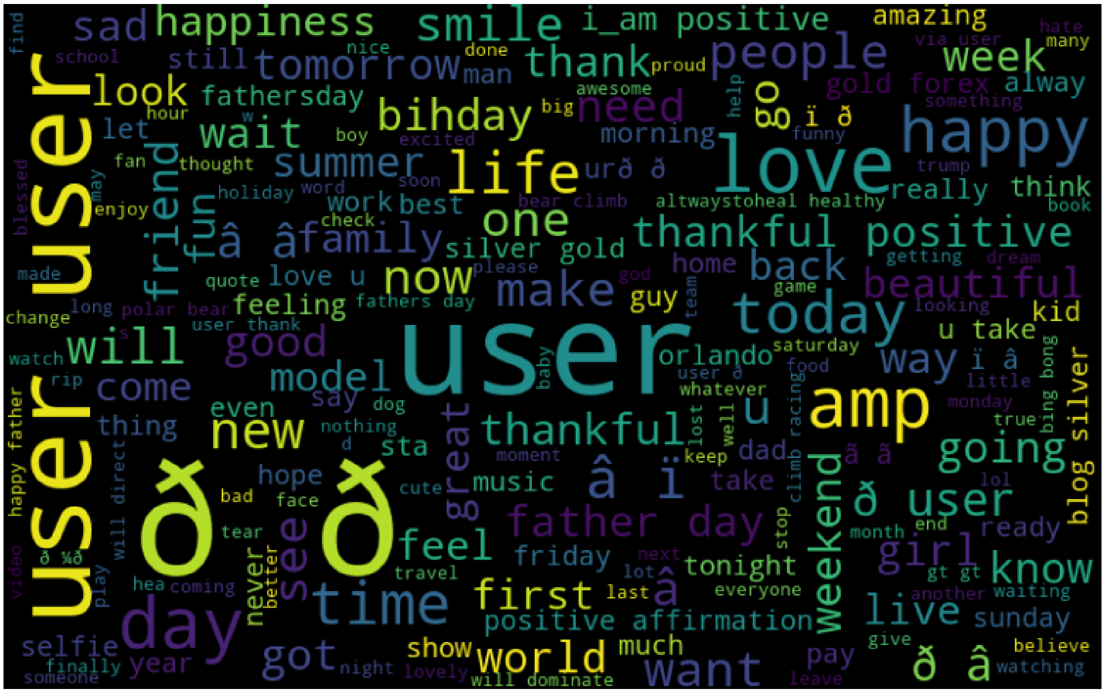


```
[15]: import matplotlib.pyplot as plt
      from wordcloud import WordCloud

      wordcloud = WordCloud(background_color='white', width=1000, height=1000).
       ↪generate_from_frequencies(dict(words_freq))

      plt.figure(figsize=(10, 8))
      plt.imshow(wordcloud)
      plt.title("WordCloud - Vocabulary from Reviews", fontsize=22)
      plt.axis('off')
      plt.show()
```



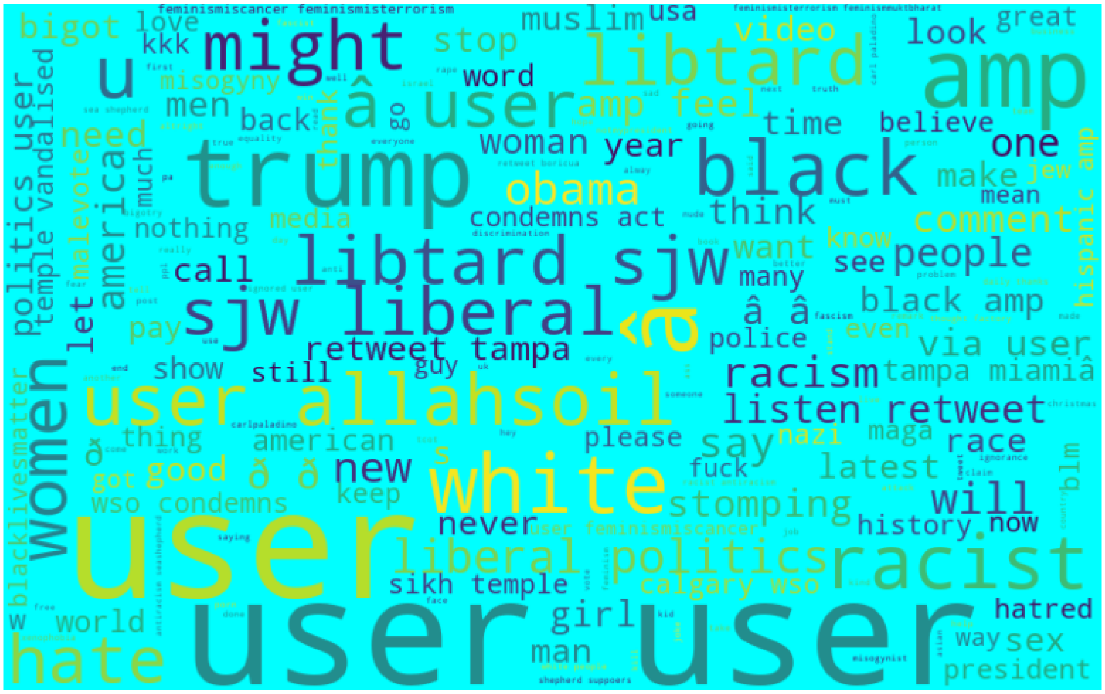
[16]:
```python
normal_words =' '.join([text for text in train['tweet'][train['label'] == 0]])

wordcloud = WordCloud(width=800, height=500, random_state = 0, max_font_size = 110).generate(normal_words)
plt.figure(figsize=(10, 7))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.title('The Neutral Words')
plt.show()
```

The Neutral Words



[17]:
```python
negative_words =' '.join([text for text in train['tweet'][train['label'] == 1]])

wordcloud = WordCloud(background_color = 'cyan', width=800, height=500, random_state = 0, max_font_size = 110).generate(negative_words)
plt.figure(figsize=(10, 7))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.title('The Negative Words')
plt.show()
```

The Negative Words

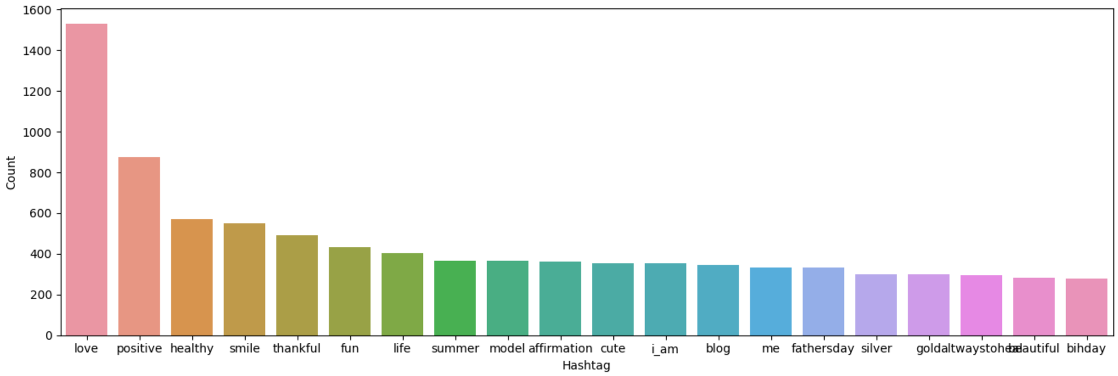


```
[18]: # collecting the hashtags
import re

def hashtag_extract(x):
    hashtags = []

    for i in x:
        ht = re.findall(r"#(\w+)", i)
        hashtags.append(ht)

    return hashtags
```

```
[19]: # extracting hashtags from non racist/sexist tweets
HT_regular = hashtag_extract(train['tweet'][train['label'] == 0])

# extracting hashtags from racist/sexist tweets
HT_negative = hashtag_extract(train['tweet'][train['label'] == 1])

# unnesting list
HT_regular = sum(HT_regular,[])
HT_negative = sum(HT_negative,[])
```

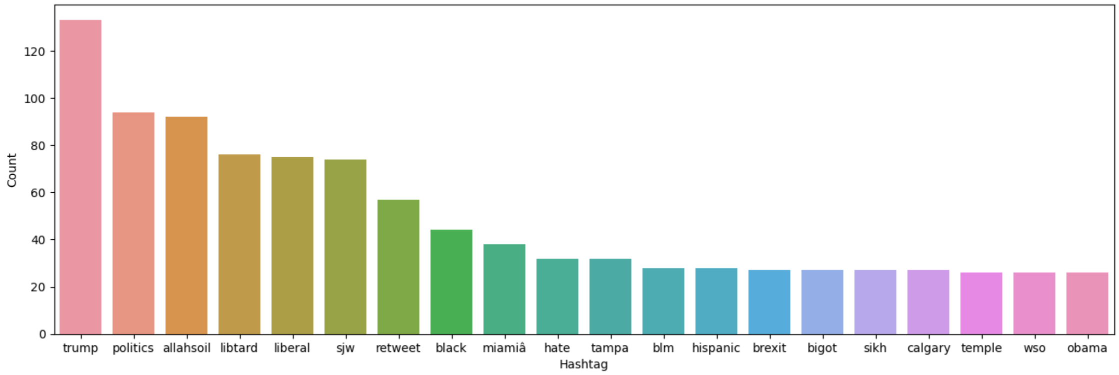
[20]:

```
import nltk
from nltk import FreqDist
import pandas as pd
a = nltk.FreqDist(HT_regular)
d = pd.DataFrame({'Hashtag': list(a.keys()),
                  'Count': list(a.values())})

# selecting top 20 most frequent hashtags
d = d.nlargest(columns="Count", n = 20)
plt.figure(figsize=(16,5))
ax = sns.barplot(data=d, x= "Hashtag", y = "Count")
ax.set(ylabel = 'Count')
plt.show()
```



[21]:

```
a = nltk.FreqDist(HT_negative)
d = pd.DataFrame({'Hashtag': list(a.keys()),
                  'Count': list(a.values())})

# selecting top 20 most frequent hashtags
d = d.nlargest(columns="Count", n = 20)
plt.figure(figsize=(16,5))
ax = sns.barplot(data=d, x= "Hashtag", y = "Count")
ax.set(ylabel = 'Count')
plt.show()
```



[4]:

```
!pip install transformers
from transformers import AutoTokenizer
from transformers import AutoModelForSequenceClassification
from scipy.special import softmax
```

```
Collecting transformers
  Obtaining dependency information for transformers from https://files.pythonhos
ted.org/packages/21/02/ae8e595f45b6c8edee07913892b3b41f5f5f273962ad98851dc6a564b
bb9/transformers-4.31.0-py3-none-any.whl.metadata
  Downloading transformers-4.31.0-py3-none-any.whl.metadata (116 kB)
     ---------------------------------------- 0.0/116.9 kB ? eta -:--:--
     ---------- ----------------------------- 30.7/116.9 kB ? eta -:--:--
     ------------ ------------------------- 41.0/116.9 kB 393.8 kB/s eta 0:00:01
     -------------------- ----------------- 61.4/116.9 kB 550.5 kB/s eta 0:00:01
     ------------------------------ ------- 92.2/116.9 kB 525.1 kB/s eta 0:00:01
     ------------------------------------ 116.9/116.9 kB 525.5 kB/s eta 0:00:00
Requirement already satisfied: filelock in c:\users\prashant
priyadarshi\appdata\local\programs\python\python310\lib\site-packages (from
transformers) (3.12.2)
Collecting huggingface-hub<1.0,>=0.14.1 (from transformers)
  Obtaining dependency information for huggingface-hub<1.0,>=0.14.1 from https:/
/files.pythonhosted.org/packages/7f/c4/adcbe9a696c135578cabcbdd7331332daad4d49b7
c43688bc2d36b3a47d2/huggingface_hub-0.16.4-py3-none-any.whl.metadata
  Downloading huggingface_hub-0.16.4-py3-none-any.whl.metadata (12 kB)
Requirement already satisfied: numpy>=1.17 in c:\users\prashant
priyadarshi\appdata\local\programs\python\python310\lib\site-packages (from
transformers) (1.24.2)
Requirement already satisfied: packaging>=20.0 in c:\users\prashant
priyadarshi\appdata\local\programs\python\python310\lib\site-packages (from
transformers) (23.0)
Requirement already satisfied: pyyaml>=5.1 in c:\users\prashant
priyadarshi\appdata\local\programs\python\python310\lib\site-packages (from
transformers) (6.0)
Requirement already satisfied: regex!=2019.12.17 in c:\users\prashant
```

```
priyadarshi\appdata\local\programs\python\python310\lib\site-packages (from
transformers) (2023.6.3)
Requirement already satisfied: requests in c:\users\prashant
priyadarshi\appdata\local\programs\python\python310\lib\site-packages (from
transformers) (2.31.0)
Collecting tokenizers!=0.11.3,<0.14,>=0.11.1 (from transformers)
  Downloading tokenizers-0.13.3-cp310-cp310-win_amd64.whl (3.5 MB)
     ---------------------------------------- 0.0/3.5 MB ? eta -:--:--
     ---------------------------------------- 0.0/3.5 MB ? eta -:--:--
     ---------------------------------------- 0.0/3.5 MB ? eta -:--:--
     - -------------------------------------- 0.1/3.5 MB 3.3 MB/s eta 0:00:02
     - -------------------------------------- 0.1/3.5 MB 2.4 MB/s eta 0:00:02
     - -------------------------------------- 0.2/3.5 MB 1.2 MB/s eta 0:00:03
     -- ------------------------------------- 0.2/3.5 MB 1.1 MB/s eta 0:00:04
     -- ------------------------------------- 0.2/3.5 MB 939.4 kB/s eta 0:00:04
     -- ------------------------------------- 0.2/3.5 MB 888.8 kB/s eta 0:00:04
     --- ------------------------------------ 0.3/3.5 MB 853.3 kB/s eta 0:00:04
     --- ------------------------------------ 0.3/3.5 MB 874.1 kB/s eta 0:00:04
     ---- ----------------------------------- 0.3/3.5 MB 866.5 kB/s eta 0:00:04
     ---- ----------------------------------- 0.4/3.5 MB 859.0 kB/s eta 0:00:04
     ----- ---------------------------------- 0.4/3.5 MB 860.2 kB/s eta 0:00:04
     ----- ---------------------------------- 0.5/3.5 MB 906.4 kB/s eta 0:00:04
     ------ --------------------------------- 0.5/3.5 MB 863.4 kB/s eta 0:00:04
     ------ --------------------------------- 0.6/3.5 MB 950.3 kB/s eta 0:00:04
     ------- -------------------------------- 0.7/3.5 MB 990.5 kB/s eta 0:00:03
     -------- ------------------------------- 0.8/3.5 MB 1.1 MB/s eta 0:00:03
     --------- ------------------------------ 0.8/3.5 MB 1.1 MB/s eta 0:00:03
     ---------- ----------------------------- 0.9/3.5 MB 1.1 MB/s eta 0:00:03
     ----------- ---------------------------- 1.0/3.5 MB 1.1 MB/s eta 0:00:03
     ------------ --------------------------- 1.1/3.5 MB 1.2 MB/s eta 0:00:03
     ------------- -------------------------- 1.2/3.5 MB 1.2 MB/s eta 0:00:02
     -------------- ------------------------- 1.3/3.5 MB 1.3 MB/s eta 0:00:02
     ---------------- ----------------------- 1.4/3.5 MB 1.3 MB/s eta 0:00:02
     ----------------- ---------------------- 1.5/3.5 MB 1.3 MB/s eta 0:00:02
     ------------------ --------------------- 1.6/3.5 MB 1.4 MB/s eta 0:00:02
     ------------------- -------------------- 1.7/3.5 MB 1.5 MB/s eta 0:00:02
     --------------------- ------------------ 1.8/3.5 MB 1.5 MB/s eta 0:00:02
     --------------------- ------------------ 1.9/3.5 MB 1.5 MB/s eta 0:00:02
     ----------------------- ---------------- 2.0/3.5 MB 1.5 MB/s eta 0:00:01
     ------------------------ --------------- 2.2/3.5 MB 1.6 MB/s eta 0:00:01
     ------------------------- -------------- 2.2/3.5 MB 1.6 MB/s eta 0:00:01
     -------------------------- ------------- 2.3/3.5 MB 1.6 MB/s eta 0:00:01
     ---------------------------- ----------- 2.5/3.5 MB 1.6 MB/s eta 0:00:01
     ----------------------------- ---------- 2.6/3.5 MB 1.7 MB/s eta 0:00:01
     ------------------------------- -------- 2.8/3.5 MB 1.7 MB/s eta 0:00:01
     -------------------------------- ------- 2.8/3.5 MB 1.7 MB/s eta 0:00:01
     ---------------------------------- ----- 2.9/3.5 MB 1.7 MB/s eta 0:00:01
     ----------------------------------- ---- 3.1/3.5 MB 1.8 MB/s eta 0:00:01
     ------------------------------------ --- 3.2/3.5 MB 1.8 MB/s eta 0:00:01
     -------------------------------------- - 3.3/3.5 MB 1.8 MB/s eta 0:00:01
     -------------------------------------- - 3.3/3.5 MB 1.8 MB/s eta 0:00:01
     -------------------------------------- - 3.4/3.5 MB 1.7 MB/s eta 0:00:01
     ---------------------------------------  3.5/3.5 MB 1.7 MB/s eta 0:00:01
     ---------------------------------------- 3.5/3.5 MB 1.7 MB/s eta 0:00:00
Collecting safetensors>=0.3.1 (from transformers)
  Downloading safetensors-0.3.1-cp310-cp310-win_amd64.whl (263 kB)
     ---------------------------------------- 0.0/263.7 kB ? eta -:--:--
     ------------ -------------------------- 81.9/263.7 kB 4.8 MB/s eta 0:00:01
     -------------------- ----------------- 143.4/263.7 kB 1.7 MB/s eta 0:00:01
     ----------------------------------- -- 245.8/263.7 kB 1.7 MB/s eta 0:00:01
     -------------------------------------- 263.7/263.7 kB 1.6 MB/s eta 0:00:00
Requirement already satisfied: tqdm>=4.27 in c:\users\prashant
priyadarshi\appdata\local\programs\python\python310\lib\site-packages (from
transformers) (4.65.0)
Collecting fsspec (from huggingface-hub<1.0,>=0.14.1->transformers)
  Obtaining dependency information for fsspec from https://files.pythonhosted.or
g/packages/e3/bd/4c0a4619494188a9db5d77e2100ab7d544a42e76b2447869d8e124e981d8/fs
spec-2023.6.0-py3-none-any.whl.metadata
  Downloading fsspec-2023.6.0-py3-none-any.whl.metadata (6.7 kB)
Requirement already satisfied: typing-extensions>=3.7.4.3 in c:\users\prashant
priyadarshi\appdata\local\programs\python\python310\lib\site-packages (from
huggingface-hub<1.0,>=0.14.1->transformers) (4.6.2)
Requirement already satisfied: colorama in c:\users\prashant
priyadarshi\appdata\local\programs\python\python310\lib\site-packages (from
tqdm>=4.27->transformers) (0.4.4)
Requirement already satisfied: charset-normalizer<4,>=2 in c:\users\prashant
priyadarshi\appdata\local\programs\python\python310\lib\site-packages (from
requests->transformers) (3.2.0)
Requirement already satisfied: idna<4,>=2.5 in c:\users\prashant
priyadarshi\appdata\local\programs\python\python310\lib\site-packages (from
requests->transformers) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in c:\users\prashant
priyadarshi\appdata\local\programs\python\python310\lib\site-packages (from
requests->transformers) (2.0.3)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\prashant
priyadarshi\appdata\local\programs\python\python310\lib\site-packages (from
requests->transformers) (2023.5.7)
Downloading transformers-4.31.0-py3-none-any.whl (7.4 MB)
   ---------------------------------------- 0.0/7.4 MB ? eta -:--:--
   ---------------------------------------- 0.1/7.4 MB 2.6 MB/s eta 0:00:03
    --------------------------------------- 0.1/7.4 MB 1.7 MB/s eta 0:00:05
   - -------------------------------------- 0.2/7.4 MB 1.5 MB/s eta 0:00:05
   - -------------------------------------- 0.2/7.4 MB 1.5 MB/s eta 0:00:05
   - -------------------------------------- 0.3/7.4 MB 1.2 MB/s eta 0:00:06
   - -------------------------------------- 0.4/7.4 MB 1.3 MB/s eta 0:00:06
   -- ------------------------------------- 0.4/7.4 MB 1.4 MB/s eta 0:00:06
-- ------------------------------------- 0.5/7.4 MB 1.4 MB/s eta 0:00:06
-- ------------------------------------- 0.5/7.4 MB 1.3 MB/s eta 0:00:06
--- ------------------------------------ 0.6/7.4 MB 1.3 MB/s eta 0:00:06
--- ------------------------------------ 0.6/7.4 MB 1.3 MB/s eta 0:00:06
--- ------------------------------------ 0.7/7.4 MB 1.3 MB/s eta 0:00:06
---- ----------------------------------- 0.8/7.4 MB 1.3 MB/s eta 0:00:06
---- ----------------------------------- 0.8/7.4 MB 1.3 MB/s eta 0:00:06
---- ----------------------------------- 0.9/7.4 MB 1.3 MB/s eta 0:00:05
----- ---------------------------------- 1.0/7.4 MB 1.3 MB/s eta 0:00:05
----- ---------------------------------- 1.0/7.4 MB 1.3 MB/s eta 0:00:05
----- ---------------------------------- 1.1/7.4 MB 1.3 MB/s eta 0:00:05
------ --------------------------------- 1.2/7.4 MB 1.3 MB/s eta 0:00:05
------ --------------------------------- 1.2/7.4 MB 1.3 MB/s eta 0:00:05
------ --------------------------------- 1.2/7.4 MB 1.3 MB/s eta 0:00:05
------- -------------------------------- 1.3/7.4 MB 1.3 MB/s eta 0:00:05
------- -------------------------------- 1.4/7.4 MB 1.3 MB/s eta 0:00:05
------- -------------------------------- 1.4/7.4 MB 1.3 MB/s eta 0:00:05
------- -------------------------------- 1.5/7.4 MB 1.3 MB/s eta 0:00:05
-------- ------------------------------- 1.5/7.4 MB 1.3 MB/s eta 0:00:05
-------- ------------------------------- 1.6/7.4 MB 1.3 MB/s eta 0:00:05
-------- ------------------------------- 1.6/7.4 MB 1.3 MB/s eta 0:00:05
-------- ------------------------------- 1.6/7.4 MB 1.2 MB/s eta 0:00:05
-------- ------------------------------- 1.6/7.4 MB 1.2 MB/s eta 0:00:05
--------- ------------------------------ 1.8/7.4 MB 1.2 MB/s eta 0:00:05
---------- ----------------------------- 2.0/7.4 MB 1.3 MB/s eta 0:00:05
----------- ---------------------------- 2.1/7.4 MB 1.3 MB/s eta 0:00:04
----------- ---------------------------- 2.1/7.4 MB 1.3 MB/s eta 0:00:04
----------- ---------------------------- 2.2/7.4 MB 1.3 MB/s eta 0:00:04
------------ --------------------------- 2.3/7.4 MB 1.4 MB/s eta 0:00:04
------------ --------------------------- 2.3/7.4 MB 1.3 MB/s eta 0:00:04
------------ --------------------------- 2.3/7.4 MB 1.3 MB/s eta 0:00:04
------------ --------------------------- 2.3/7.4 MB 1.3 MB/s eta 0:00:04
------------ --------------------------- 2.4/7.4 MB 1.3 MB/s eta 0:00:04
------------- -------------------------- 2.4/7.4 MB 1.3 MB/s eta 0:00:04
------------- -------------------------- 2.5/7.4 MB 1.3 MB/s eta 0:00:04
------------- -------------------------- 2.5/7.4 MB 1.2 MB/s eta 0:00:04
------------- -------------------------- 2.5/7.4 MB 1.2 MB/s eta 0:00:04
-------------- ------------------------- 2.6/7.4 MB 1.2 MB/s eta 0:00:04
-------------- ------------------------- 2.7/7.4 MB 1.2 MB/s eta 0:00:04
-------------- ------------------------- 2.7/7.4 MB 1.2 MB/s eta 0:00:04
-------------- ------------------------- 2.7/7.4 MB 1.2 MB/s eta 0:00:04
-------------- ------------------------- 2.7/7.4 MB 1.2 MB/s eta 0:00:04
-------------- ------------------------- 2.8/7.4 MB 1.2 MB/s eta 0:00:04
--------------- ------------------------ 2.8/7.4 MB 1.2 MB/s eta 0:00:04
--------------- ------------------------ 2.8/7.4 MB 1.2 MB/s eta 0:00:04
--------------- ------------------------ 2.9/7.4 MB 1.2 MB/s eta 0:00:04
--------------- ------------------------ 2.9/7.4 MB 1.2 MB/s eta 0:00:04
--------------- ------------------------ 2.9/7.4 MB 1.1 MB/s eta 0:00:04
```

```
--------------- ------------------------ 2.9/7.4 MB 1.1 MB/s eta 0:00:04
--------------- ------------------------ 2.9/7.4 MB 1.1 MB/s eta 0:00:04
---------------- ----------------------- 3.0/7.4 MB 1.1 MB/s eta 0:00:04
---------------- ----------------------- 3.0/7.4 MB 1.1 MB/s eta 0:00:04
---------------- ----------------------- 3.0/7.4 MB 1.1 MB/s eta 0:00:04
---------------- ----------------------- 3.1/7.4 MB 1.1 MB/s eta 0:00:05
---------------- ----------------------- 3.1/7.4 MB 1.1 MB/s eta 0:00:05
---------------- ----------------------- 3.1/7.4 MB 1.1 MB/s eta 0:00:04
----------------- ---------------------- 3.1/7.4 MB 1.1 MB/s eta 0:00:04
----------------- ---------------------- 3.2/7.4 MB 1.1 MB/s eta 0:00:04
----------------- ---------------------- 3.2/7.4 MB 1.1 MB/s eta 0:00:04
----------------- ---------------------- 3.2/7.4 MB 1.0 MB/s eta 0:00:05
----------------- ---------------------- 3.2/7.4 MB 1.0 MB/s eta 0:00:05
----------------- ---------------------- 3.3/7.4 MB 1.0 MB/s eta 0:00:05
----------------- ---------------------- 3.3/7.4 MB 1.0 MB/s eta 0:00:05
----------------- ---------------------- 3.3/7.4 MB 1.0 MB/s eta 0:00:05
------------------ --------------------- 3.3/7.4 MB 1.0 MB/s eta 0:00:05
------------------ --------------------- 3.4/7.4 MB 992.1 kB/s eta 0:00:05
------------------ --------------------- 3.4/7.4 MB 987.5 kB/s eta 0:00:05
------------------ --------------------- 3.4/7.4 MB 980.0 kB/s eta 0:00:05
------------------ --------------------- 3.4/7.4 MB 975.7 kB/s eta 0:00:05
------------------ --------------------- 3.5/7.4 MB 968.5 kB/s eta 0:00:05
------------------ --------------------- 3.5/7.4 MB 965.9 kB/s eta 0:00:05
------------------- -------------------- 3.5/7.4 MB 957.9 kB/s eta 0:00:05
------------------- -------------------- 3.5/7.4 MB 948.5 kB/s eta 0:00:05
------------------- -------------------- 3.5/7.4 MB 942.1 kB/s eta 0:00:05
------------------- -------------------- 3.6/7.4 MB 942.5 kB/s eta 0:00:05
------------------- -------------------- 3.6/7.4 MB 935.0 kB/s eta 0:00:05
------------------- -------------------- 3.6/7.4 MB 938.1 kB/s eta 0:00:04
------------------- -------------------- 3.7/7.4 MB 933.5 kB/s eta 0:00:04
-------------------- ------------------- 3.7/7.4 MB 930.4 kB/s eta 0:00:04
-------------------- ------------------- 3.7/7.4 MB 930.7 kB/s eta 0:00:04
-------------------- ------------------- 3.8/7.4 MB 927.5 kB/s eta 0:00:04
-------------------- ------------------- 3.8/7.4 MB 927.1 kB/s eta 0:00:04
-------------------- ------------------- 3.9/7.4 MB 922.9 kB/s eta 0:00:04
--------------------- ------------------ 3.9/7.4 MB 925.8 kB/s eta 0:00:04
--------------------- ------------------ 3.9/7.4 MB 925.2 kB/s eta 0:00:04
--------------------- ------------------ 4.0/7.4 MB 923.6 kB/s eta 0:00:04
--------------------- ------------------ 4.0/7.4 MB 923.6 kB/s eta 0:00:04
--------------------- ------------------ 4.0/7.4 MB 916.5 kB/s eta 0:00:04
---------------------- ----------------- 4.1/7.4 MB 919.8 kB/s eta 0:00:04
---------------------- ----------------- 4.1/7.4 MB 917.0 kB/s eta 0:00:04
---------------------- ----------------- 4.2/7.4 MB 918.8 kB/s eta 0:00:04
---------------------- ----------------- 4.2/7.4 MB 920.6 kB/s eta 0:00:04
----------------------- ---------------- 4.3/7.4 MB 924.2 kB/s eta 0:00:04
----------------------- ---------------- 4.3/7.4 MB 922.7 kB/s eta 0:00:04
----------------------- ---------------- 4.4/7.4 MB 924.4 kB/s eta 0:00:04
------------------------ --------------- 4.5/7.4 MB 934.6 kB/s eta 0:00:04
```

```
   ------------------------ --------------- 4.5/7.4 MB 942.5 kB/s eta 0:00:04
   ------------------------- -------------- 4.6/7.4 MB 951.5 kB/s eta 0:00:03
   ------------------------- -------------- 4.8/7.4 MB 966.4 kB/s eta 0:00:03
   -------------------------- ------------- 4.9/7.4 MB 980.0 kB/s eta 0:00:03
   -------------------------- ------------- 5.0/7.4 MB 989.2 kB/s eta 0:00:03
   --------------------------- ------------ 5.0/7.4 MB 996.2 kB/s eta 0:00:03
   --------------------------- ------------ 5.1/7.4 MB 994.2 kB/s eta 0:00:03
   --------------------------- ------------ 5.1/7.4 MB 993.9 kB/s eta 0:00:03
   ---------------------------- ----------- 5.2/7.4 MB 994.8 kB/s eta 0:00:03
   ---------------------------- ----------- 5.2/7.4 MB 999.7 kB/s eta 0:00:03
   ---------------------------- ----------- 5.3/7.4 MB 999.5 kB/s eta 0:00:03
   ---------------------------- ----------- 5.3/7.4 MB 992.6 kB/s eta 0:00:03
   ----------------------------- ---------- 5.4/7.4 MB 996.4 kB/s eta 0:00:03
   ----------------------------- ---------- 5.5/7.4 MB 1.0 MB/s eta 0:00:02
   ------------------------------ --------- 5.6/7.4 MB 1.0 MB/s eta 0:00:02
   ------------------------------ --------- 5.7/7.4 MB 1.0 MB/s eta 0:00:02
   ------------------------------- -------- 5.8/7.4 MB 1.0 MB/s eta 0:00:02
   ------------------------------- -------- 5.9/7.4 MB 1.0 MB/s eta 0:00:02
   -------------------------------- ------- 6.0/7.4 MB 1.1 MB/s eta 0:00:02
   -------------------------------- ------- 6.0/7.4 MB 1.1 MB/s eta 0:00:02
   -------------------------------- ------- 6.1/7.4 MB 1.1 MB/s eta 0:00:02
   --------------------------------- ------ 6.1/7.4 MB 1.1 MB/s eta 0:00:02
   --------------------------------- ------ 6.1/7.4 MB 1.1 MB/s eta 0:00:02
   --------------------------------- ------ 6.1/7.4 MB 1.1 MB/s eta 0:00:02
   --------------------------------- ------ 6.3/7.4 MB 1.1 MB/s eta 0:00:02
   ---------------------------------- ----- 6.3/7.4 MB 1.1 MB/s eta 0:00:02
   ---------------------------------- ----- 6.4/7.4 MB 1.1 MB/s eta 0:00:01
   ----------------------------------- ---- 6.5/7.4 MB 1.1 MB/s eta 0:00:01
   ----------------------------------- ---- 6.5/7.4 MB 1.1 MB/s eta 0:00:01
   ----------------------------------- ---- 6.6/7.4 MB 1.1 MB/s eta 0:00:01
   ------------------------------------ --- 6.8/7.4 MB 1.1 MB/s eta 0:00:01
   ------------------------------------- -- 6.9/7.4 MB 1.1 MB/s eta 0:00:01
   ------------------------------------- -- 7.0/7.4 MB 1.1 MB/s eta 0:00:01
   -------------------------------------- - 7.1/7.4 MB 1.1 MB/s eta 0:00:01
   -------------------------------------- - 7.2/7.4 MB 1.1 MB/s eta 0:00:01
   ---------------------------------------  7.3/7.4 MB 1.1 MB/s eta 0:00:01
   ---------------------------------------  7.4/7.4 MB 1.1 MB/s eta 0:00:01
   ---------------------------------------- 7.4/7.4 MB 1.1 MB/s eta 0:00:00
Downloading huggingface_hub-0.16.4-py3-none-any.whl (268 kB)
   ---------------------------------------- 0.0/268.8 kB ? eta -:--:--
   ---------------------------------------- 0.0/268.8 kB ? eta -:--:--
   ---------------------------------------- 0.0/268.8 kB ? eta -:--:--
   ---------------------------------------  266.2/268.8 kB 8.3 MB/s eta 0:00:01
   ---------------------------------------- 268.8/268.8 kB 5.5 MB/s eta 0:00:00
Downloading fsspec-2023.6.0-py3-none-any.whl (163 kB)
   ---------------------------------------- 0.0/163.8 kB ? eta -:--:--
   --------------------------- ------------ 112.6/163.8 kB 3.2 MB/s eta 0:00:01
   ---------------------------------------- 163.8/163.8 kB 1.9 MB/s eta 0:00:00
```

```
Installing collected packages: tokenizers, safetensors, fsspec, huggingface-hub,
transformers
Successfully installed fsspec-2023.6.0 huggingface-hub-0.16.4 safetensors-0.3.1
tokenizers-0.13.3 transformers-4.31.0
```

[19]:

```
# Tokenizing the words present in the training set
tokenized_tweet = train['tweet'].apply(lambda x: x.split())

# Importing gensim
import gensim

# Converting tokenized tweets to list of sentences
sentences = tokenized_tweet.tolist()

# Creating a word to vector model
model_w2v = gensim.models.Word2Vec(
    sentences,
    vector_size=200,  # Desired number of features/independent variables
    window=5,  # Context window size
    min_count=2,
    sg=1,  # 1 for skip-gram model
    hs=0,
    negative=10,  # For negative sampling
    workers=2,  # Number of cores
    seed=34
)

model_w2v.train(sentences, total_examples=len(sentences), epochs=20)
```

[19]:

```
(6109121, 8411580)
```

[20]:

```
model_w2v.wv.most_similar(positive = "dinner")
```

[20]:

```
[('spaghetti', 0.646221399307251),
 ('#prosecco', 0.604263961315155),
 ('coaching', 0.600287139415741),
 ('#wanderlust', 0.5991036891937256),
 ('podium', 0.5844422578811646),
 ('fluffy', 0.5742541551589966),
 ('7!', 0.5731244683265686),
 ('pampered', 0.5727431178092957),
 ('sister!!', 0.5724947452545166),
 ('snuggle', 0.572417140007019)]
```

[21]:

```
model_w2v.wv.most_similar(positive = "cancer")
```

```
[21]: [('champion,', 0.7131642699241638),
        ('level.', 0.6987432837486267),
        ('ways.', 0.6946672797203064),
        ('tolerance', 0.6924739480018616),
        ('ownership', 0.6918320655822754),
        ('roots', 0.689439594745636),
        ('#merica', 0.6882787346839905),
        ('weapon', 0.6857432723045349),
        ('aol', 0.6841249465942383),
        ('#guncontrolplease', 0.6771395802497864)]
```

```
[ ]:
```