Importing Libraries

In [3]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import re
import string
%pip install nltk
import nltk
import warnings
%matplotlib inline
```

```
Collecting nltkNote: you may need to restart the kernel to use updated packages.

  Downloading nltk-3.9.1-py3-none-any.whl.metadata (2.9 kB)
Requirement already satisfied: click in c:\users\gauta\anaconda3\lib\site-packages (from nltk) (8.1.7)
Collecting joblib (from nltk)
  Downloading joblib-1.4.2-py3-none-any.whl.metadata (5.4 kB)
Collecting regex>=2021.8.3 (from nltk)
  Downloading regex-2024.7.24-cp311-cp311-win_amd64.whl.metadata (41 kB)
     ---------------------------------------- 0.0/41.5 kB ? eta -:--:--
     ------------------ ------------------- 20.5/41.5 kB 320.0 kB/s eta 0:00:01
     -------------------------------------  41.0/41.5 kB 388.9 kB/s eta 0:00:01
     -------------------------------------  41.0/41.5 kB 388.9 kB/s eta 0:00:01
     -------------------------------------- 41.5/41.5 kB 199.4 kB/s eta 0:00:00
Requirement already satisfied: tqdm in c:\users\gauta\anaconda3\lib\site-packages (from nltk) (4.65.0)
Requirement already satisfied: colorama in c:\users\gauta\anaconda3\lib\site-packages (from click->nltk) (0.4.6)
Downloading nltk-3.9.1-py3-none-any.whl (1.5 MB)
   ---------------------------------------- 0.0/1.5 MB ? eta -:--:--
   --- ------------------------------------ 0.1/1.5 MB 2.4 MB/s eta 0:00:01
   ------------ --------------------------- 0.5/1.5 MB 5.3 MB/s eta 0:00:01
   ------------------------------------- -- 1.4/1.5 MB 9.9 MB/s eta 0:00:01
   ---------------------------------------  1.5/1.5 MB 9.6 MB/s eta 0:00:01
   ---------------------------------------- 1.5/1.5 MB 6.9 MB/s eta 0:00:00
Downloading regex-2024.7.24-cp311-cp311-win_amd64.whl (269 kB)
   ---------------------------------------- 0.0/269.7 kB ? eta -:--:--
   ---------------------------------------  266.2/269.7 kB 17.1 MB/s eta 0:00:01
   ---------------------------------------- 269.7/269.7 kB 4.2 MB/s eta 0:00:00
Downloading joblib-1.4.2-py3-none-any.whl (301 kB)
   ---------------------------------------- 0.0/301.8 kB ? eta -:--:--
   ---------------------------------------  297.0/301.8 kB ? eta -:--:--
   ---------------------------------------- 301.8/301.8 kB 3.7 MB/s eta 0:00:00
Installing collected packages: regex, joblib, nltk
Successfully installed joblib-1.4.2 nltk-3.9.1 regex-2024.7.24
```

Importing Dataset

In [4]:

```
df = pd.read_csv(r"C:\Users\gauta\OneDrive\Documents\brainwave intern\Twitter Sentiments.csv")
df.head()
```

Out[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

Preprocessing

In [6]:

```
df.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
```

Data cleaning

In [8]:

```
def remove_pattern(input_txt, pattern):
    r = re.findall(pattern, input_txt)
    for word in r:
        input_txt = re.sub(word, "", input_txt)
    return input_txt


df.head()
```

Out[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

In [9]:

```
df['clean_tweet'] = np.vectorize(remove_pattern)(df['tweet'], "@[\w]*")


df.head()
```

Out[9]:

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

In [10]:

```
df['clean_tweet'] = df['clean_tweet'].str.replace("[^a-zA-Z#]", " ")


df.head()
```

Out[10]:

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

In [11]:

```
df['clean_tweet'] = df['clean_tweet'].apply(lambda x: " ".join([w for w in x.split() if len(w)>3]))


df.head()
```

Out[11]:

	id	label	tweet	clean_tweet
0	1	0	@user when a father is dysfunctional and is s...	when father dysfunctional selfish drags kids i...
1	2	0	@user @user thanks for #lyft credit i can't us...	thanks #lyft credit can't cause they don't off...
2	3	0	bihday your majesty	bihday your majesty
3	4	0	#model i love u take with u all the time in ...	#model love take with time urð±!!! ððð...
4	5	0	factsguide: society now #motivation	factsguide: society #motivation

In [12]:

```
tokenized_tweet = df['clean_tweet'].apply(lambda x: x.split())


tokenized_tweet.head()
```

Out[12]:

```
0    [when, father, dysfunctional, selfish, drags, ...
1    [thanks, #lyft, credit, can't, cause, they, do...
2                              [bihday, your, majesty]
3    [#model, love, take, with, time, urð±!!!, ð...
4                  [factsguide:, society, #motivation]
Name: clean_tweet, dtype: object
```

In [13]:

```
from nltk.stem.porter import PorterStemmer
stemmer = PorterStemmer()

tokenized_tweet = tokenized_tweet.apply(lambda sentence: [stemmer.stem(word) for word in sentence])

tokenized_tweet.head()
```

Out[13]:

```
0    [when, father, dysfunct, selfish, drag, kid, i...
1    [thank, #lyft, credit, can't, caus, they, don'...
2                              [bihday, your, majesti]
3    [#model, love, take, with, time, urð±!!!, ð...
4                        [factsguide:, societi, #motiv]
Name: clean_tweet, dtype: object
```

In [14]:

```
for i in range(len(tokenized_tweet)):
    tokenized_tweet[i] = " ".join(tokenized_tweet[i])

df['clean_tweet'] = tokenized_tweet

df.head()
```

Out[14]:

	id	label	tweet	clean_tweet
0	1	0	@user when a father is dysfunctional and is s...	when father dysfunct selfish drag kid into dys...
1	2	0	@user @user thanks for #lyft credit i can't us...	thank #lyft credit can't caus they don't offer...
2	3	0	bihday your majesty	bihday your majesti
3	4	0	#model i love u take with u all the time in ...	#model love take with time urð±!!! ððð...
4	5	0	factsguide: society now #motivation	factsguide: societi #motiv

Exploratory Data Analysis (EDA)

In [15]:

```
!pip install wordcloud
```

```
Collecting wordcloud
  Downloading wordcloud-1.9.3-cp311-cp311-win_amd64.whl.metadata (3.5 kB)
Requirement already satisfied: numpy>=1.6.1 in c:\users\gauta\anaconda3\lib\site-packages (from wordcloud) (1.26
.4)
Requirement already satisfied: pillow in c:\users\gauta\anaconda3\lib\site-packages (from wordcloud) (10.2.0)
Requirement already satisfied: matplotlib in c:\users\gauta\anaconda3\lib\site-packages (from wordcloud) (3.9.0)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\gauta\anaconda3\lib\site-packages (from matplotlib->
wordcloud) (1.2.0)
Requirement already satisfied: cycler>=0.10 in c:\users\gauta\anaconda3\lib\site-packages (from matplotlib->word
cloud) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\gauta\anaconda3\lib\site-packages (from matplotlib-
>wordcloud) (4.53.0)
Requirement already satisfied: kiwisolver>=1.3.1 in c:\users\gauta\anaconda3\lib\site-packages (from matplotlib-
>wordcloud) (1.4.5)
Requirement already satisfied: packaging>=20.0 in c:\users\gauta\anaconda3\lib\site-packages (from matplotlib->w
ordcloud) (23.1)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\gauta\anaconda3\lib\site-packages (from matplotlib->
wordcloud) (3.1.2)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\gauta\anaconda3\lib\site-packages (from matplotl
ib->wordcloud) (2.8.2)
Requirement already satisfied: six>=1.5 in c:\users\gauta\anaconda3\lib\site-packages (from python-dateutil>=2.7
->matplotlib->wordcloud) (1.16.0)
Downloading wordcloud-1.9.3-cp311-cp311-win_amd64.whl (300 kB)
   ---------------------------------------- 0.0/300.2 kB ? eta -:--:--
   ---------------------------------------- 0.0/300.2 kB ? eta -:--:--
   - -------------------------------------- 10.2/300.2 kB ? eta -:--:--
   - -------------------------------------- 10.2/300.2 kB ? eta -:--:--
   ----- ---------------------------------- 41.0/300.2 kB 281.8 kB/s eta 0:00:01
   -------------- ----------------------- 112.6/300.2 kB 595.3 kB/s eta 0:00:01
   ---------------------------------------  297.0/300.2 kB 1.4 MB/s eta 0:00:01
   ---------------------------------------- 300.2/300.2 kB 1.0 MB/s eta 0:00:00
Installing collected packages: wordcloud
Successfully installed wordcloud-1.9.3
```

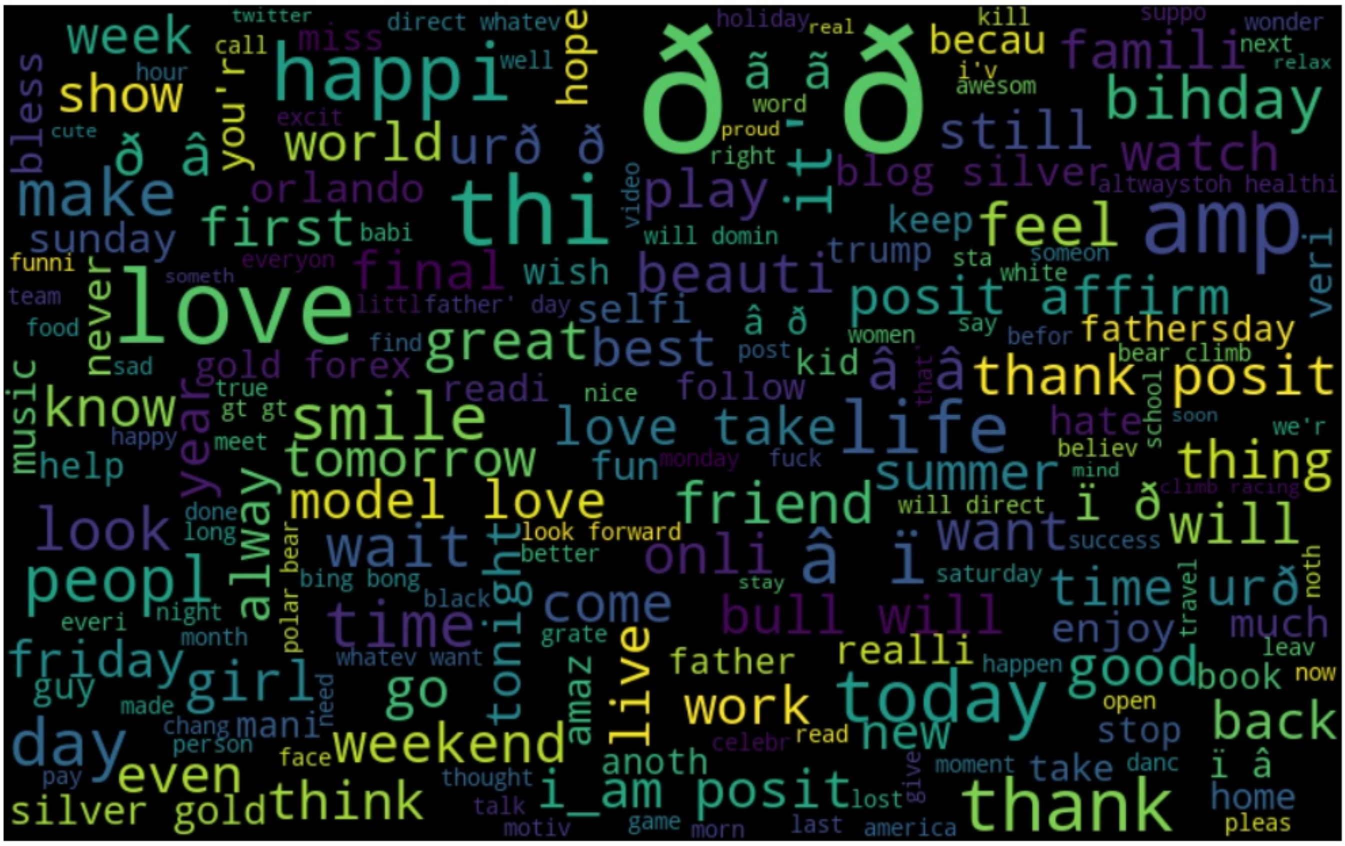
In [16]:

```
all_words = " ".join([sentence for sentence in df['clean_tweet']])

from wordcloud import WordCloud
wordcloud = WordCloud(width=800, height=500, random_state=42, max_font_size=100).generate(all_words)

# plot the graph
plt.figure(figsize=(15,8))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')

plt.show()
```

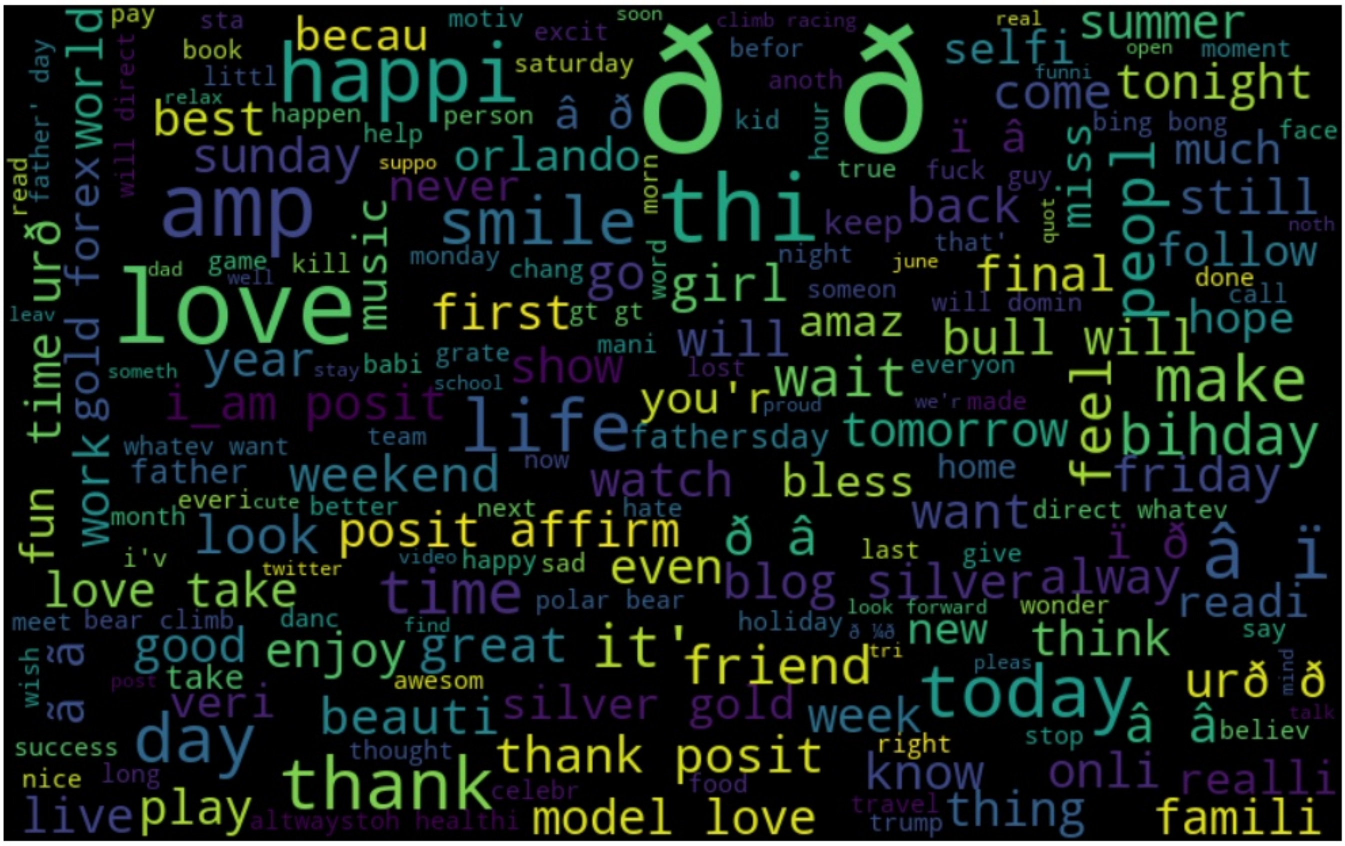


```
In [32]: all_words = " ".join([sentence for sentence in df['clean_tweet'][df['label']==0]])

         wordcloud = WordCloud(width=800, height=500, random_state=42, max_font_size=100, colormap='viridis').generate(a

         # plot the graph
         plt.figure(figsize=(15,8))
         plt.imshow(wordcloud, interpolation='bilinear')
         plt.axis('off')

         plt.show()
```

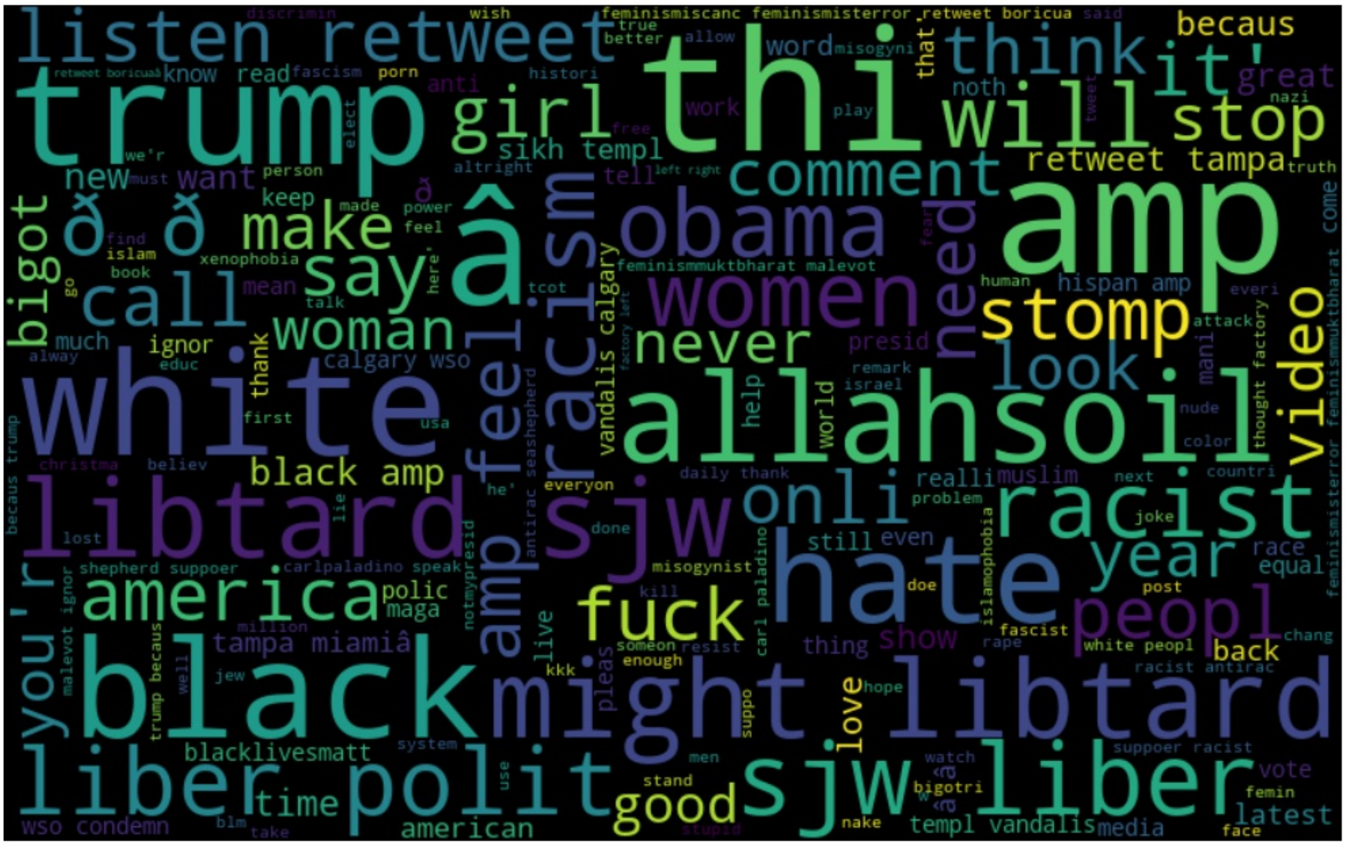


```
In [18]: all_words = " ".join([sentence for sentence in df['clean_tweet'][df['label']==1]])

         wordcloud = WordCloud(width=800, height=500, random_state=42, max_font_size=100).generate(all_words)
```

```
# plot the graph
plt.figure(figsize=(15,8))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')

plt.show()
```



In [19]:
```
def hashtag_extract(tweets):
    hashtags = []

    # loop words in the tweet
    for tweet in tweets:
        ht = re.findall(r"#(\w+)", tweet)
        hashtags.append(ht)
    return hashtags
```

In [20]:
```
ht_positive = hashtag_extract(df['clean_tweet'][df['label']==0])


ht_negative = hashtag_extract(df['clean_tweet'][df['label']==1])


ht_positive[:5]
```

Out[20]:
```
[['run'], ['lyft', 'disapoint', 'getthank'], [], ['model'], ['motiv']]
```

In [ ]:
```
ht_positive = sum(ht_positive, [])
ht_negative = sum(ht_negative, [])
```

In [24]:
```
ht_positive[:5]
```

Out[24]:
```
['run', 'lyft', 'disapoint', 'getthank', 'model']
```

In [25]:
```
freq = nltk.FreqDist(ht_positive)
d = pd.DataFrame({'Hashtag': list(freq.keys()),
                  'Count': list(freq.values())})


d.head()
```

Out[25]:

	Hashtag	Count
0	run	70
1	lyft	2
2	disapoint	1
3	getthank	2
4	model	374

In [31]:

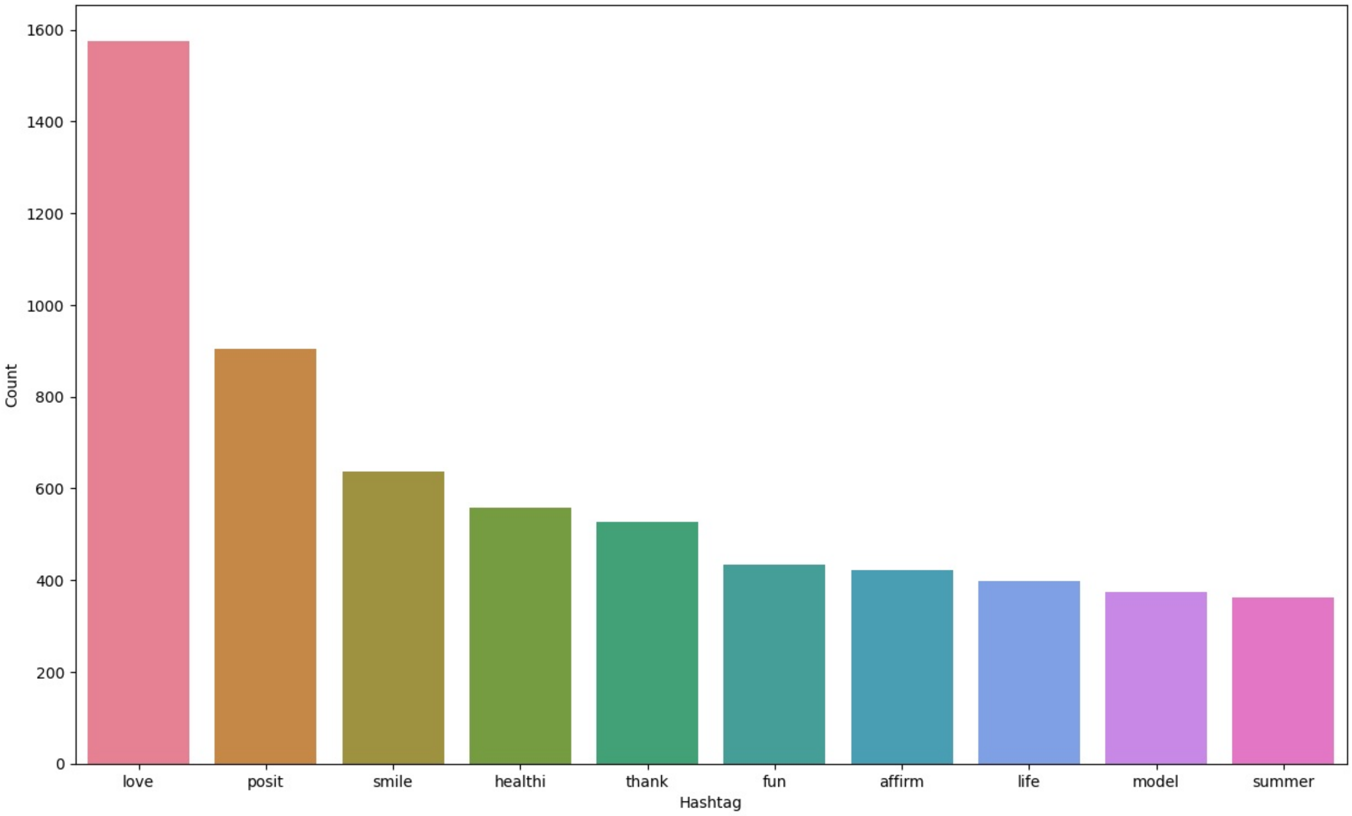
```
d = d.nlargest(columns='Count', n=10)
plt.figure(figsize=(15,9))
sns.barplot(data=d, x='Hashtag', y='Count', palette="husl")

plt.show()
```

```
C:\Users\gauta\AppData\Local\Temp\ipykernel_8768\1646232964.py:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable
to `hue` and set `legend=False` for the same effect.

  sns.barplot(data=d, x='Hashtag', y='Count', palette="husl")
```



In [33]:

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


d.head()
```

Out[33]:

	Hashtag	Count
0	cnn	9
1	michigan	2
2	tcot	14
3	australia	6
4	opkillingbay	2

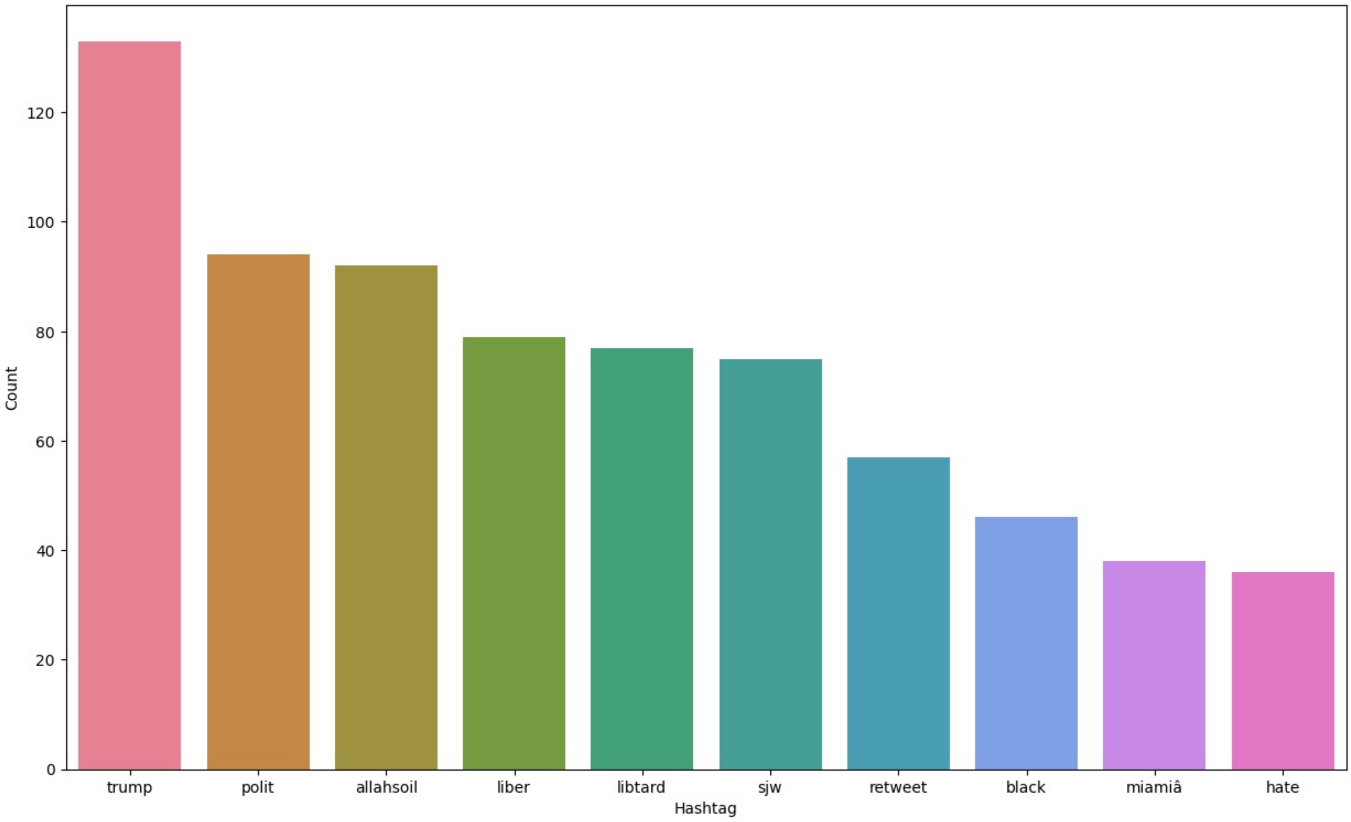
In [39]:

```
d = d.nlargest(columns='Count', n=10)
plt.figure(figsize=(15,9))
sns.barplot(data=d, x='Hashtag', y='Count', palette="husl")
plt.show()
```

```
C:\Users\gauta\AppData\Local\Temp\ipykernel_8768\3490887480.py:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable
to `hue` and set `legend=False` for the same effect.

  sns.barplot(data=d, x='Hashtag', y='Count', palette="husl")
```



Input split

```
In [41]: pip install scikit-learn
```

```
Collecting scikit-learn
  Downloading scikit_learn-1.5.1-cp311-cp311-win_amd64.whl.metadata (12 kB)
Requirement already satisfied: numpy>=1.19.5 in c:\users\gauta\anaconda3\lib\site-packages (from scikit-learn) (
1.26.4)
Collecting scipy>=1.6.0 (from scikit-learn)
  Downloading scipy-1.14.1-cp311-cp311-win_amd64.whl.metadata (60 kB)
     ---------------------------------------- 0.0/60.8 kB ? eta -:--:--
     ------ --------------------------------- 10.2/60.8 kB ? eta -:--:--
     ------------ ------------------------- 20.5/60.8 kB 165.2 kB/s eta 0:00:01
     ------------------------- ------------ 41.0/60.8 kB 281.8 kB/s eta 0:00:01
     -------------------------------- ----- 51.2/60.8 kB 327.7 kB/s eta 0:00:01
     -------------------------------------- 60.8/60.8 kB 249.0 kB/s eta 0:00:00
Requirement already satisfied: joblib>=1.2.0 in c:\users\gauta\anaconda3\lib\site-packages (from scikit-learn) (
1.4.2)
Collecting threadpoolctl>=3.1.0 (from scikit-learn)
  Downloading threadpoolctl-3.5.0-py3-none-any.whl.metadata (13 kB)
Downloading scikit_learn-1.5.1-cp311-cp311-win_amd64.whl (11.0 MB)
   ---------------------------------------- 0.0/11.0 MB ? eta -:--:--
    --------------------------------------- 0.1/11.0 MB 4.3 MB/s eta 0:00:03
   - -------------------------------------- 0.5/11.0 MB 5.2 MB/s eta 0:00:03
   ---- ----------------------------------- 1.2/11.0 MB 9.2 MB/s eta 0:00:02
   -------- ------------------------------- 2.2/11.0 MB 11.9 MB/s eta 0:00:01
   ---------- ----------------------------- 2.8/11.0 MB 12.0 MB/s eta 0:00:01
   ----------- ---------------------------- 3.1/11.0 MB 11.5 MB/s eta 0:00:01
   ------------ --------------------------- 3.3/11.0 MB 11.1 MB/s eta 0:00:01
   ------------- -------------------------- 3.6/11.0 MB 10.0 MB/s eta 0:00:01
   -------------- ------------------------- 3.9/11.0 MB 9.9 MB/s eta 0:00:01
   -------------- ------------------------- 3.9/11.0 MB 9.9 MB/s eta 0:00:01
   ---------------- ----------------------- 4.5/11.0 MB 8.9 MB/s eta 0:00:01
   ----------------- ---------------------- 4.8/11.0 MB 8.9 MB/s eta 0:00:01
   ------------------ --------------------- 5.1/11.0 MB 8.5 MB/s eta 0:00:01
   ------------------- -------------------- 5.4/11.0 MB 8.3 MB/s eta 0:00:01
   -------------------- ------------------- 5.6/11.0 MB 8.2 MB/s eta 0:00:01
   --------------------- ------------------ 5.9/11.0 MB 8.1 MB/s eta 0:00:01
   ---------------------- ----------------- 6.2/11.0 MB 7.9 MB/s eta 0:00:01
   ----------------------- ---------------- 6.5/11.0 MB 7.9 MB/s eta 0:00:01
   ------------------------ --------------- 6.8/11.0 MB 7.8 MB/s eta 0:00:01
   ------------------------- -------------- 7.1/11.0 MB 7.7 MB/s eta 0:00:01
   -------------------------- ------------- 7.4/11.0 MB 7.6 MB/s eta 0:00:01
   ------------------------------------ --- 40.5/44.8 MB 6.1 MB/s eta 0:00:01
   ------------------------------------ --- 40.8/44.8 MB 6.2 MB/s eta 0:00:01
   ------------------------------------ --- 41.1/44.8 MB 6.1 MB/s eta 0:00:01
   ------------------------------------ --- 41.4/44.8 MB 6.1 MB/s eta 0:00:01
   ------------------------------------- -- 41.7/44.8 MB 6.3 MB/s eta 0:00:01
   ------------------------------------- -- 42.0/44.8 MB 6.1 MB/s eta 0:00:01
   ------------------------------------- -- 42.3/44.8 MB 6.2 MB/s eta 0:00:01
   ------------------------------------- -- 42.5/44.8 MB 6.1 MB/s eta 0:00:01
   -------------------------------------- - 42.8/44.8 MB 6.1 MB/s eta 0:00:01
   -------------------------------------- - 43.1/44.8 MB 6.2 MB/s eta 0:00:01
   -------------------------------------- - 43.4/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  43.7/44.8 MB 6.1 MB/s eta 0:00:01
   ---------------------------------------  44.0/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.3/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.6/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------  44.8/44.8 MB 6.2 MB/s eta 0:00:01
   ---------------------------------------- 44.8/44.8 MB 3.4 MB/s eta 0:00:00
Downloading threadpoolctl-3.5.0-py3-none-any.whl (18 kB)
Installing collected packages: threadpoolctl, scipy, scikit-learn
Successfully installed scikit-learn-1.5.1 scipy-1.14.1 threadpoolctl-3.5.0
Note: you may need to restart the kernel to use updated packages.
```

In [42]:

```
from sklearn.feature_extraction.text import CountVectorizer
bow_vectorizer = CountVectorizer(max_df=0.90, min_df=2, max_features=1000, stop_words='english')
bow = bow_vectorizer.fit_transform(df['clean_tweet'])
```

In [43]:

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(bow, df['label'], random_state=42, test_size=0.25)
```

Training the model

In [44]:

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import f1_score, accuracy_score
```

In [45]:

```
model = LogisticRegression()
model.fit(x_train, y_train)
```

Out[45]:

```
▾ LogisticRegression ⓘ ?

LogisticRegression()
```

In [46]:

```
pred = model.predict(x_test)
f1_score(y_test, pred)
```

Out[46]:

```
0.506508875739645
```

In [47]:

```
accuracy_score(y_test,pred)
```

Out[47]:

```
0.9478162933299963
```

In [48]:

```
pred_prob = model.predict_proba(x_test)
pred = pred_prob[:, 1] >= 0.3
pred = pred.astype(np.int64)

f1_score(y_test, pred)
```

```
Out[48]: 0.5575992255566312
```

```
In [49]: accuracy_score(y_test,pred)
```

```
Out[49]: 0.9428106619947441
```

```
In [50]: pred_prob[0][1] >= 0.3
```

```
Out[50]: False
```

```
In [ ]:
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js