In [2]:
```
import re
import pandas as pd
pd.set_option("display.max_colwidth", 200)
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import string
import nltk
import warnings
warnings.filterwarnings("ignore", category=DeprecationWarning)
%matplotlib inline
```

In [3]:
```
train  = pd.read_csv('dataset/tweets_train.csv')
test = pd.read_csv('dataset/tweets_test.csv')
```

In [5]:
```
train[train['label'] == 0].head(10)
```

Out[5]:

	id	label	tweet
0	1	0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run
1	2	0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in urð+!!! ðððð¦ð¦ð¦
4	5	0	factsguide: society now #motivation
5	6	0	[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there. #allshowandnogo
6	7	0	@user camping tomorrow @user @user @user @user @user @user @user dannyâ¦
7	8	0	the next school year is the year for exams.ð¯ can't think about that ð #school #exams #hate #imagine #actorslife #revolutionschool #girl
8	9	0	we won!!! love the land!!! #allin #cavs #champions #cleveland #clevelandcavaliers â¦
9	10	0	@user @user welcome here ! i'm it's so #gr8 !

In [6]:
```
train[train['label'] == 1].head(10)
```

Out[6]:

	id	label	tweet
13	14	1	@user #cnn calls #michigan middle school 'build the wall' chant " #tcot
14	15	1	no comment! in #australia #opkillingbay #seashepherd #helpcovedolphins #thecove #helpcovedolphins
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'd need something like this. again. #neverump #xenophobia
56	57	1	@user lets fight against #love #peace

	id	label	tweet
68	69	1	ð©the white establishment can't have blk folx running around loving themselves and promoting our greatness
77	78	1	@user hey, white people: you can call people 'white' by @user #race #identity #medâ¦
82	83	1	how the #altright uses & insecurity to lure men into #whitesupremacy
111	112	1	@user i'm not interested in a #linguistics that doesn't address #race & . racism is

In [7]:
```
train.shape, test.shape
```

Out[7]:
```
((31962, 3), (17197, 2))
```

In [8]:
```
train["label"].value_counts()
```

Out[8]:
```
0    29720
1     2242
Name: label, dtype: int64
```

Data Cleaning

In [9]:
```
def remove_pattern(input_txt, pattern):
    r = re.findall(pattern, input_txt)
    for i in r:
        input_txt = re.sub(i, '', input_txt)

    return input_txt
```

1. Removing Twitter Handles (@user)

In [10]:
```
train['tidy_tweet'] = np.vectorize(remove_pattern)(train['tweet'], "@[\w]*")
train.head()
```

Out[10]:

	id	label	tweet	tidy_tweet
0	1	0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run	when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run
1	2	0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked	thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked
2	3	0	bihday your majesty	bihday your majesty
3	4	0	#model i love u take with u all the time in urð±!!! ðððð¦ð¦ð¦	#model i love u take with u all the time in urð±!!! ðððð¦ð¦ð¦
4	5	0	factsguide: society now #motivation	factsguide: society now #motivation

2. Removing Punctuations, Numbers, and Special Characters

In [11]:
```
train['tidy_tweet'] = train['tidy_tweet'].str.replace("[^a-zA-Z#]", " ")
train.head(10)
```

Out[11]:

	id	label	tweet	tidy_tweet
0	1	0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run	when a father is dysfunctional and is so selfish he drags his kids into his dysfunction #run
1	2	0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked	thanks for #lyft credit i can t use cause they don t offer wheelchair vans in pdx #disapointed #getthanked
2	3	0	bihday your majesty	bihday your majesty
3	4	0	#model i love u take with u all the time in urð±!!! ðððð¦ð¦ð¦	#model i love u take with u all the time in ur
4	5	0	factsguide: society now #motivation	factsguide society now #motivation
5	6	0	[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there. #allshowandnogo	huge fan fare and big talking before they leave chaos and pay disputes when they get there #allshowandnogo
6	7	0	@user camping tomorrow @user @user @user @user @user @user @user dannyâ¦	camping tomorrow danny
7	8	0	the next school year is the year for exams.ð can't think about that Â #school #exams #hate #imagine #actorslife #revolutionschool #girl	the next school year is the year for exams can t think about that #school #exams #hate #imagine #actorslife #revolutionschool #girl
8	9	0	we won!!! love the land!!! #allin #cavs #champions #cleveland #clevelandcavaliers â¦	we won love the land #allin #cavs #champions #cleveland #clevelandcavaliers
9	10	0	@user @user welcome here ! i'm it's so #gr8 !	welcome here i m it s so #gr

3. Removing Short Words

In [12]:
```
train['tidy_tweet'] = train['tidy_tweet'].apply(lambda x: ' '.join([w for w in
train.head()
```

Out[12]:

	id	label	tweet	tidy_tweet
0	1	0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run	when father dysfunctional selfish drags kids into dysfunction #run
1	2	0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked	thanks #lyft credit cause they offer wheelchair vans #disapointed #getthanked
2	3	0	bihday your majesty	bihday your majesty
3	4	0	#model i love u take with u all the time in urð±!!! ðððð¦ð¦ð¦	#model love take with time
4	5	0	factsguide: society now #motivation	factsguide society #motivation

In [13]:
```
tokenized_tweet = train['tidy_tweet'].apply(lambda x: x.split()) #tokenizatio
```

In [14]:
```
tokenized_tweet.head()
```

Out[14]:
```
0                [when, father, dysfunctional, selfish, drags, kids, into, dys
function, #run]
```

```
1    [thanks, #lyft, credit, cause, they, offer, wheelchair, vans, #disapointe
d, #getthanked]
2                                                                    [bihday,
your, majesty]
3                                                                [#model, love, ta
ke, with, time]
4                                                              [factsguide, societ
y, #motivation]
Name: tidy_tweet, dtype: object
```

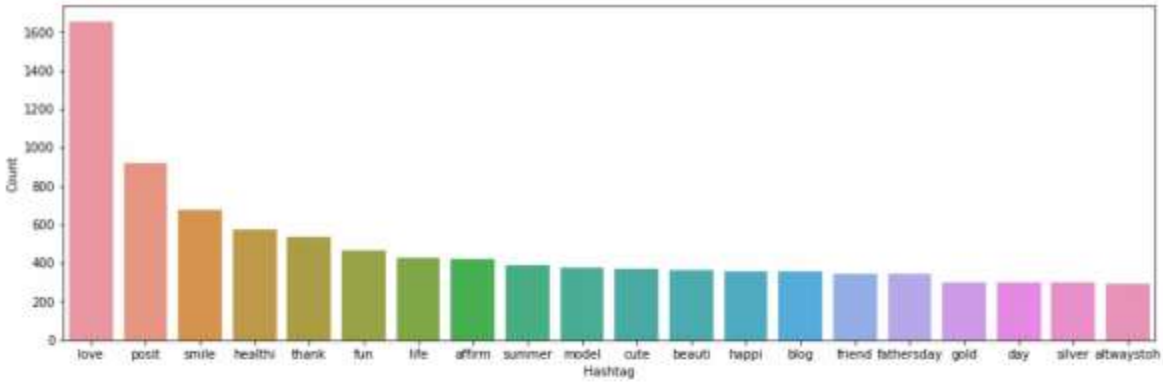
In [15]:
```
from nltk.stem.porter import *
stemmer = PorterStemmer()

tokenized_tweet = tokenized_tweet.apply(lambda x: [stemmer.stem(i) for i in x
```

In [16]:
```
for i in range(len(tokenized_tweet)):
    tokenized_tweet[i] = ' '.join(tokenized_tweet[i])

train['tidy_tweet'] = tokenized_tweet
```

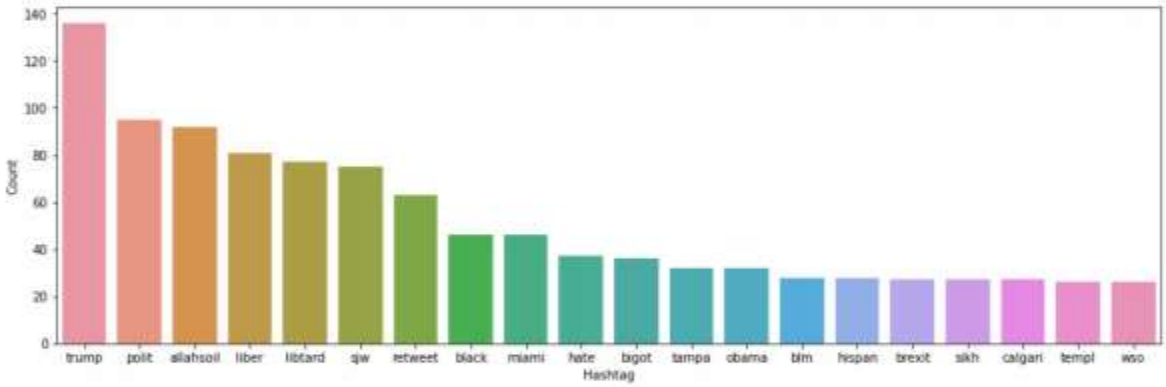
In [17]:
```
def hashtag_extract(x):
    hashtags = []
    for i in x:
        ht = re.findall(r"#(\w+)", i)
        hashtags.append(ht)

    return hashtags
```

In [18]:
```
# extracting hashtags from normal tweets
HT_regular = hashtag_extract(train['tidy_tweet'][train['label'] == 0])

# extracting hashtags from hate tweets tweets
HT_negative = hashtag_extract(train['tidy_tweet'][train['label'] == 1])

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

In [19]:
```
# Non Hate Tweets
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()
```



In [20]:
```
# Hate tweets

b = nltk.FreqDist(HT_negative)
e = pd.DataFrame({'Hashtag': list(b.keys()), 'Count': list(b.values())})

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



In [21]:
```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import f1_score
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [22]:
```
X = train["tidy_tweet"]
y = train["label"]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, ran
```

In [23]:
```
vectorizer = TfidfVectorizer()
train_vectors = vectorizer.fit_transform(X_train)
test_vectors = vectorizer.transform(X_test)
print(train_vectors.shape, test_vectors.shape)
```

```
(21414, 23783) (10548, 23783)
```

In [24]:
```
lreg = LogisticRegression()
lreg.fit(train_vectors,y_train)

from  sklearn.metrics  import accuracy_score
from sklearn import metrics

predicted = lreg.predict(test_vectors)

print("Accuracy:",accuracy_score(y_test,predicted))
print("Precision:",metrics.precision_score(y_test, predicted))
print("Recall:",metrics.recall_score(y_test, predicted))
```

```
Accuracy: 0.9489002654531665
Precision: 0.8836206896551724
Recall: 0.2859135285913529
```

In [ ]: