sentiment

May 12, 2022

[1]:
```
import numpy as np
import pandas as pd

import matplotlib.pyplot as plt
import seaborn as sns
```

[2]:
```
train = pd.read_csv('https://raw.githubusercontent.com/sharmaroshan/
↪Twitter-Sentiment-Analysis/master/train_tweet.csv')
test = pd.read_csv('https://raw.githubusercontent.com/sharmaroshan/
↪Twitter-Sentiment-Analysis/master/test_tweets.csv')
```

[3]:
```
train.head()
```

[3]:

	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

[4]:
```
test.head()
```

[4]:

	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…

[5]:
```
train.isnull().any()
test.isnull().any()
```

[5]:
```
id       False
tweet    False
dtype: bool
```

[6]:
```
train[train['label'] == 0].head(10)
```

[6]:

	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…

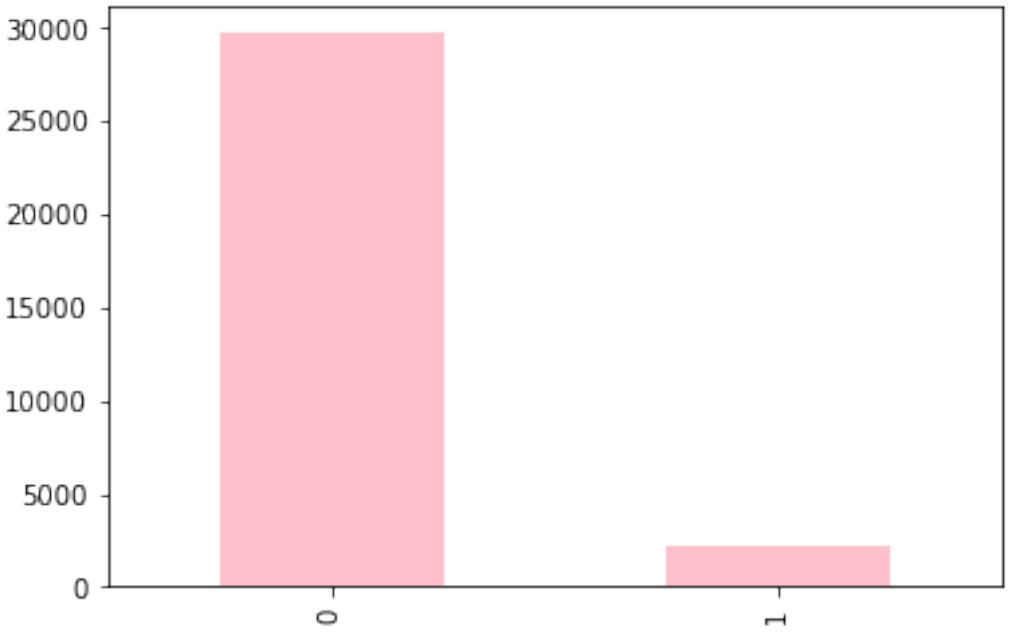
[7]:
```
train[train['label'] == 1].head(10)
```

[7]:

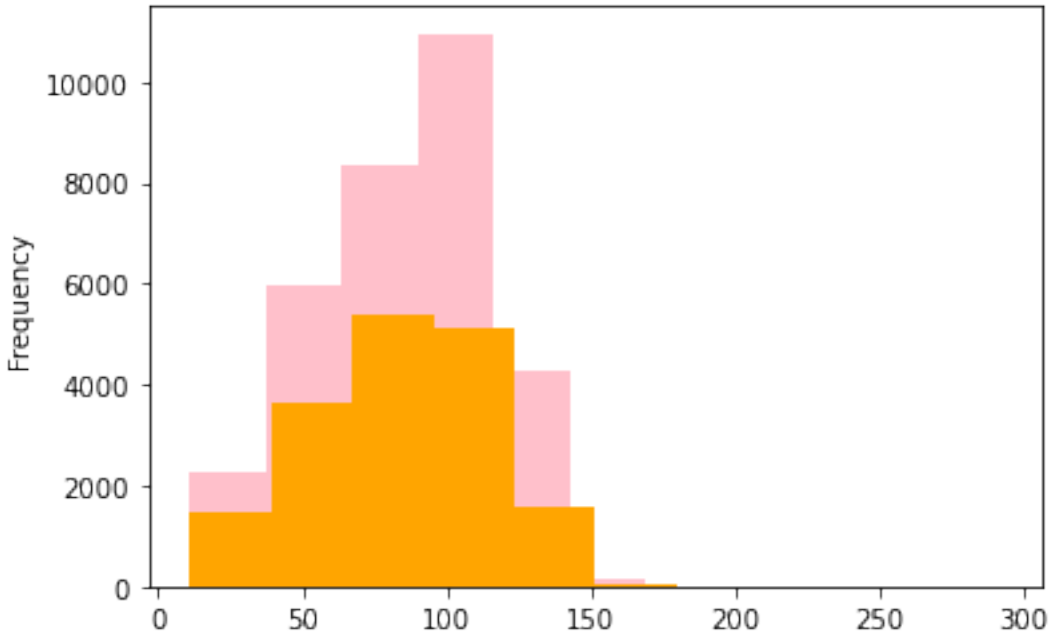
	id	label	tweet
13	14	1	@user #cnn calls #michigan middle school 'buil…
14	15	1	no comment! in #australia #opkillingbay #se…
17	18	1	retweet if you agree!
23	24	1	@user @user lumpy says i am a . prove it lumpy.
34	35	1	it's unbelievable that in the 21st century we'…
56	57	1	@user lets fight against #love #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 & insecurity to lu…
111	112	1	@user i'm not interested in a #linguistics tha…

[8]:
```
train['label'].value_counts().plot.bar(color = 'pink', figsize = (6, 4))
```

[8]: <AxesSubplot:>



[9]:
```
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))
```



[10]:
```
train['len'] = train['tweet'].str.len()
test['len'] = test['tweet'].str.len()

train.head(10)
```

[10]:

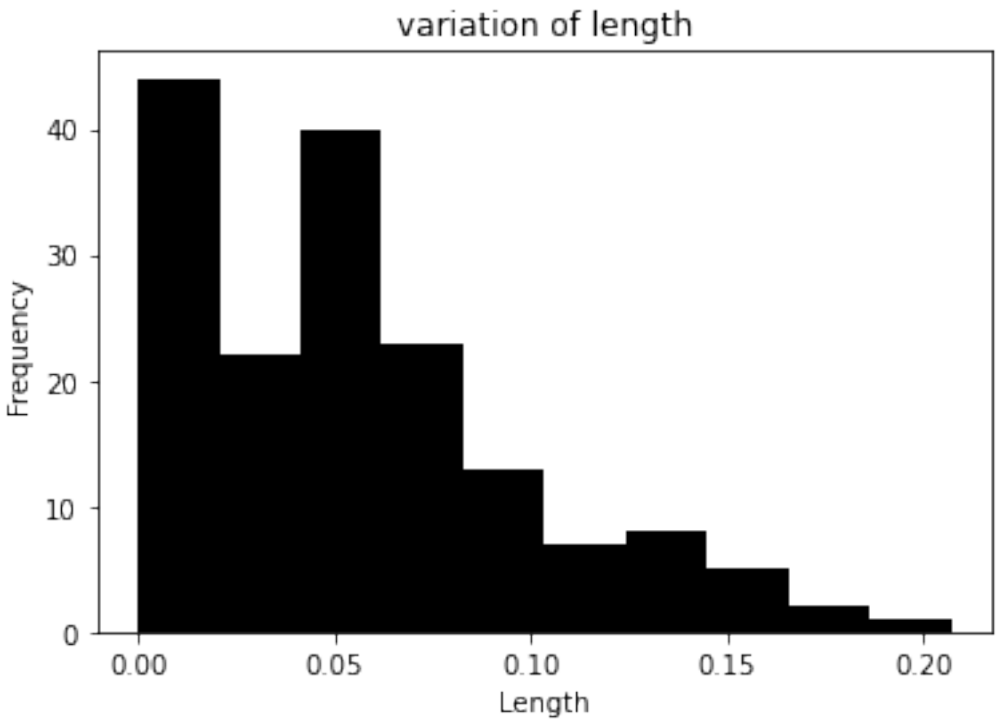
	id	label	tweet	len
0	1	0	@user when a father is dysfunctional and is s…	102
1	2	0	@user @user thanks for #lyft credit i can't us…	122
2	3	0	bihday your majesty	21
3	4	0	#model i love u take with u all the time in …	86
4	5	0	factsguide: society now #motivation	39
5	6	0	[2/2] huge fan fare and big talking 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

[11]:
```
train.groupby('label').describe()
```

[11]:

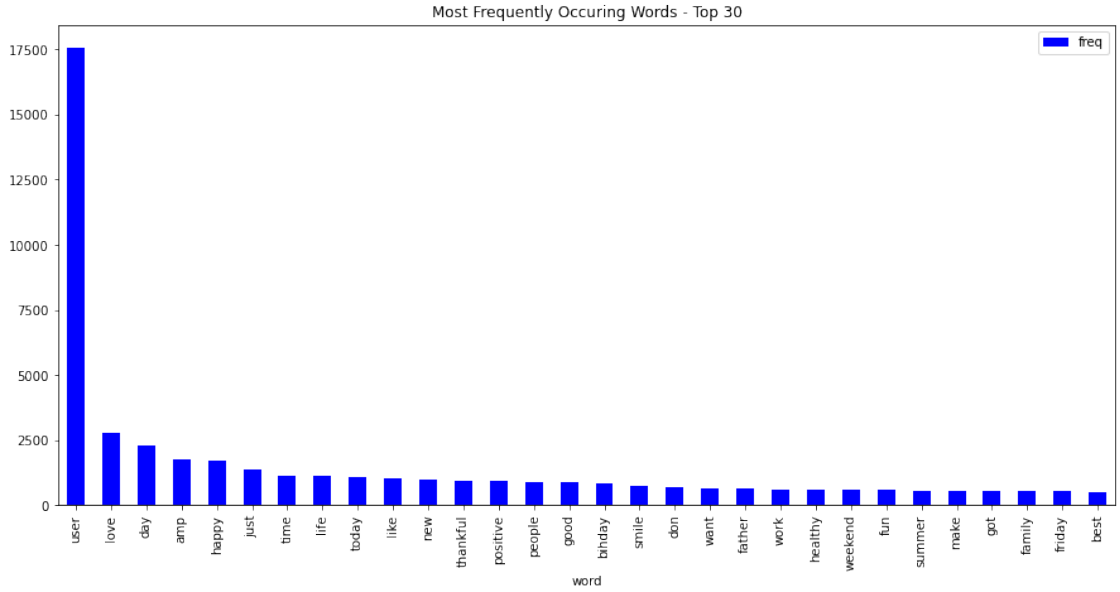
	id							len						
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
label																
0	29720.0	15974.454441	9223.783469	1.0	7981.75	15971.5	23965.25	31962.0	29720.0	84.328634	29.566484	11.0	62.0	88.0	107.0	274.0
1	2242.0	16074.896075	9267.955758	14.0	8075.25	16095.0	24022.00	31961.0	2242.0	90.187779	27.375502	12.0	69.0	96.0	111.0	152.0

[12]:
```
train.groupby('len').mean()['label'].plot.hist(color = 'black', figsize = (6,␣
↪4),)

plt.title('variation of length')
plt.xlabel('Length')
plt.show()
```



[13]:
```
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")
```

[13]:
```
Text(0.5, 1.0, 'Most Frequently Occuring Words - Top 30')
```

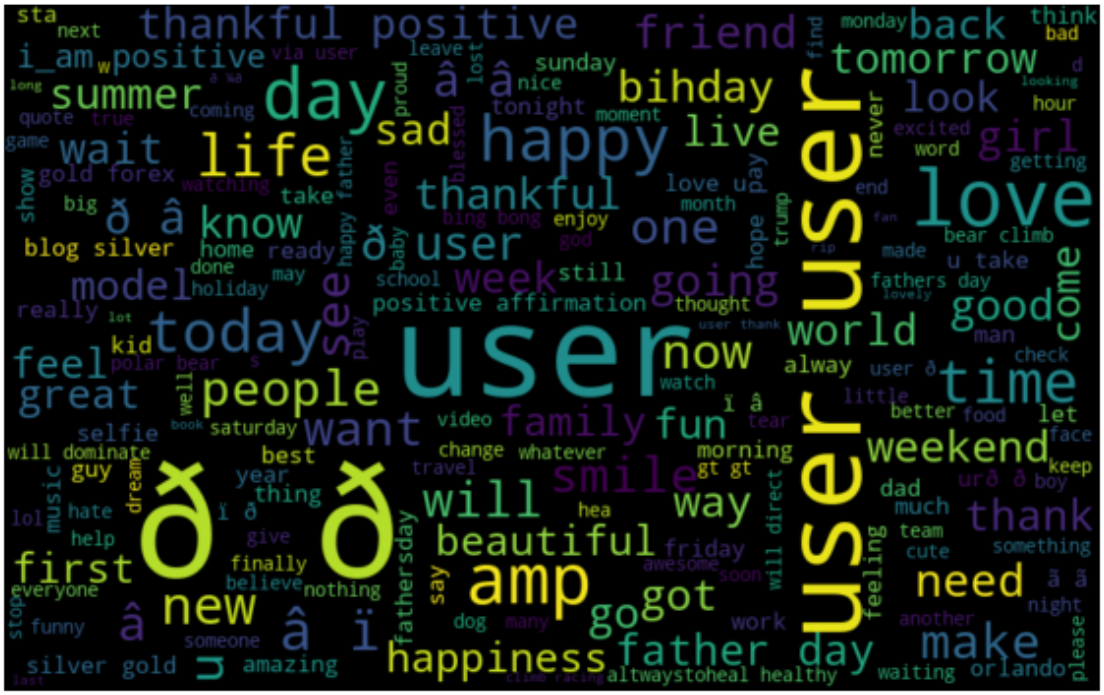


[14]:
```
from wordcloud import WordCloud

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)
```

[14]:
```
Text(0.5, 1.0, 'WordCloud - Vocabulary from Reviews')
```

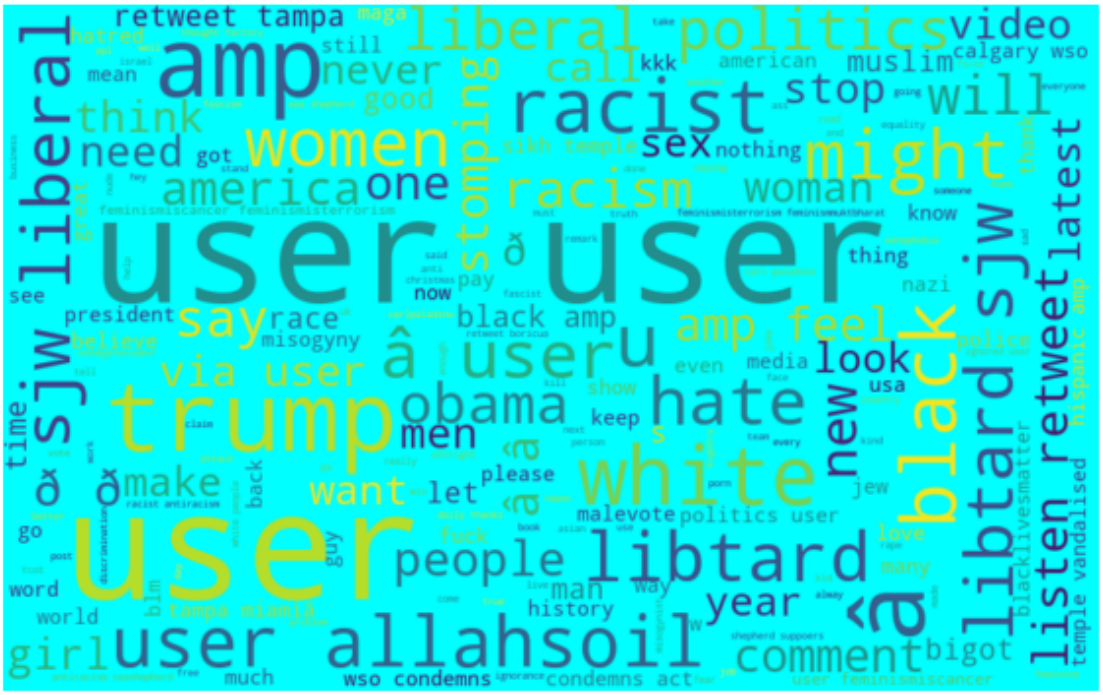
[15]:
```
normal_words =' '.join([text for text in train['tweet'][train['label'] == 0]])

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



[16]:
```
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()
```

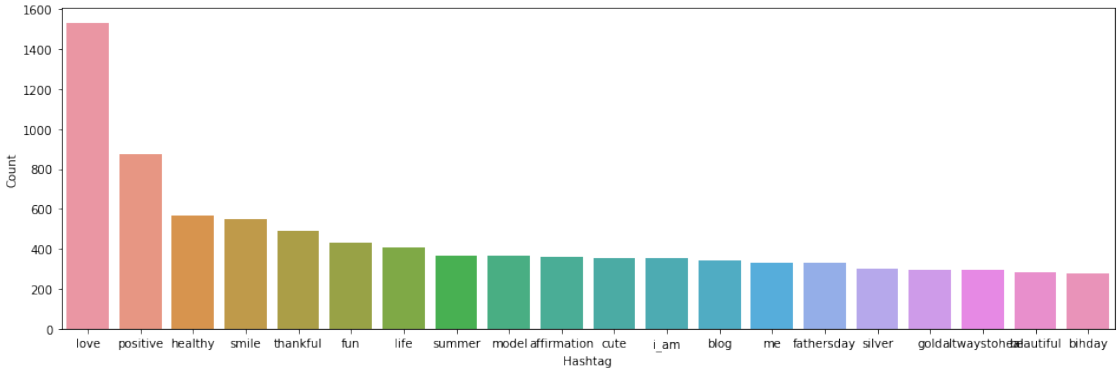
The Negative Words



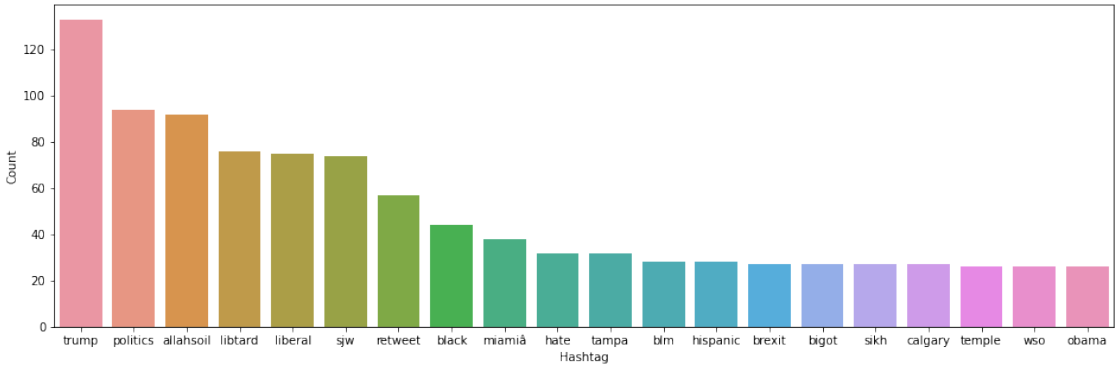
[17]:
```
def hashtag_extract(x):
    hashtags = []

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

    return hashtags
```

[20]:
```
import re


# 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,[])
```

[21]:
```
import nltk

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()
```



[22]:
```
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()
```


Count
120
100
80
60
40
20
0
trump
politics
allahsoil
libtard
liberal
sjw
retweet
black
miamiâ
hate
tampa
blm
hispanic
brexit
bigot
sikh
calgary
temple
wso
obama
Hashtag