In [2]:
```
import pandas as pd
dataset = pd.read_csv('hate_speech.csv')
dataset.head()
```

Out[2]:

	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 [3]:
```
dataset.shape
```

Out[3]:
```
(5242, 3)
```

In [5]:
```
dataset.label.value_counts()
```

Out[5]:
```
0    3000
1    2242
Name: label, dtype: int64
```

In [6]:
```
for index, tweet in enumerate(dataset["tweet"][10:15]):
    print(index+1,"_",tweet)
```

```
1 _  â#ireland consumer price index (mom) climbed from previous 0.2% to 0.5% in may   #blog #silver #gold #forex
2 _ we are so selfish. #orlando #standwithorlando #pulseshooting #orlandoshooting #biggerproblems #selfish #heabreaking   #values #love #
3 _ i get to see my daddy today!!   #80days #gettingfed
4 _ ouch...junior is angryð#got7 #junior #yugyoem   #omg
5 _ i am thankful for having a paner. #thankful #positive
```

In [8]:
```
import re
def clean_text(text):
    text = re.sub(r'[^a-zA-Z\']', ' ',text)
    text = re.sub(r'[^\x00-\x7F]+',' ', text)
    text = text.lower()
    return text
```

In [9]:
```
dataset['clean_text'] = dataset.tweet.apply(lambda x: clean_text(x))
```

In [10]:
```
dataset.head(10)
```

Out[10]:

	id	label	tweet	clean_text
0	1	0	@user when a father is dysfunctional and is s...	user when a father is dysfunctional and is s...
1	2	0	@user @user thanks for #lyft credit i can't us...	user user thanks for lyft credit i can't us...
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
5	6	0	[2/2] huge fan fare and big talking before the...	huge fan fare and big talking before the...
6	7	0	@user camping tomorrow @user @user @user @use...	user camping tomorrow user user user use...
7	8	0	the next school year is the year for exams.ð□□...	the next school year is the year for exams ...
8	9	0	we won!!! love the land!!! #allin #cavs #champ...	we won love the land allin cavs champ...
9	10	0	@user @user welcome here ! i'm it's so #gr...	user user welcome here i'm it's so gr...

In [11]:
```
from nltk.corpus import stopwords
len(stopwords.words('english'))
```

Out[11]:
```
179
```

In [13]:
```
stop = stopwords.words('english')
```

In [14]:
```
def gen_freq(text):
    word_list = []
    for tw_words in text.split():
        word_list.extend(tw_words)
    word_freq = pd.Series(word_list).value_counts()
    word_freq = word_freq.drop(stop,errors='ignore')
    return word_freq
```

In [15]:
```
def any_neg(words):
    for word in words:
        if word in ['n', 'no', 'non', 'not'] or re.search(r"\wn't", word):
            return 1
        else:
            return 0
```

In [16]:
```
def any_rare(words, rare_100):
    for word in words:
        if word in rare_100:
            return 1
        else:
            return 0
```

In [17]:
```
def is_question(words):
    for word in words:
        if word in ['when', 'what', 'how', 'why', 'who', 'where']:
            return 1
        else:
            return 0
```

In [18]:
```
word_freq = gen_freq(dataset.clean_text.str)
rare_100 = word_freq[-100:]
dataset['word_count'] = dataset.clean_text.str.split().apply(lambda x: len(x))
dataset['any_neg']=dataset.clean_text.str.split().apply(lambda x: any_neg(x))
dataset['is_question']=dataset.clean_text.str.split().apply(lambda x: is_question(x
dataset['any_rare']=dataset.clean_text.str.split().apply(lambda x: any_rare(x,rare_
dataset.clean_text.apply(lambda x: len(x))
```

Out[18]:
```
0       102
1       122
2        21
3        86
4        39
       ...
5237     59
5238     82
5239    112
5240     87
5241     67
Name: clean_text, Length: 5242, dtype: int64
```

In [ ]: