In [5]:
```
import pandas as pd
dataset=pd.read_csv('hate_speech.csv')
dataset.head()
```

Out[5]:

	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 [6]:
```
dataset.shape
```

Out[6]:
```
(5242, 3)
```

In [7]:
```
dataset.label.value_counts()
```

Out[7]:
```
0    3000
1    2242
Name: label, dtype: int64
```

In [8]:
```
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% i
n 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 [44]:
```
import re
#clean text from noise
def clean_text(text):
    #Filter to only alphabets
    text=re.sub(r'[a^-zA-Z\']',' ',text)
    text=re.sub(r'[^\x00-\x7F]+',' ',text)
    text=text.lower()
    return text
```

In [45]:
```
dataset['clean_text']=dataset.tweet.apply(lambda x:clean_text(x))
```

In [46]:
```
dataset.head(10)
```

Out[46]:

	id	label	tweet	clean_text	word_count	any_neg	is_question	any_rare	cha
0	1	0	@user when a father is dysfunctional and is s...	@ ...	3	0.0	0.0	0.0	
1	2	0	@user @user thanks for #lyft credit i can't us...	@ @ # ...	6	0.0	0.0	0.0	
2	3	0	bihday your majesty		0	NaN	NaN	NaN	
3	4	0	#model i love u take with u all the time in ...	# ...	2	0.0	0.0	0.0	
4	5	0	factsguide: society now #motivation	: #	2	0.0	0.0	0.0	
5	6	0	[2/2] huge fan fare and big talking before the...	[2/2] ...	4	0.0	0.0	0.0	
6	7	0	@user camping tomorrow @user @user @user @use...	@ @ @ @ @ ...	8	0.0	0.0	0.0	
7	8	0	the next school year is the year for exams.ð□□...	. ...	8	0.0	0.0	0.0	
8	9	0	we won!!! love the land!!! #allin #cavs #champ...	!!! !!! # # # ...	7	0.0	0.0	0.0	
9	10	0	@user @user welcome here ! i'm it's so #gr...	@ @ ! # ...	6	0.0	0.0	0.0	

In [47]:

```
from nltk.corpus import stopwords
#listening stop words
len(stopwords.words('english'))
```

Out[47]:

```
179
```

In [48]:

```
stop=stopwords.words('english')
```

In [49]:

```
def gen_freq(text):
    #will store the list of words
    word_list=[]
    #Loop over all the tweets and extract words into word_list
    for tw_words in text.split():
        word_list.extend(tw_words)
    #Create word frequencies using word_list
    word_freq=pd.Series(word_list).value_counts()
    #Drop the stopwords during the frequency calculation
    word_freq=word_freq.drop(stop_words,errors='ignore')
    return word_freq
```

In [50]:

```
#Check whether the negation term is present in the text
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 [51]:

```
#Check whether one of the 100 rare words is present in the text
def any_rare(words, rare_100):
    for word in words:
        if word in rare_100:
            return 1
        else:
            return 0
```

In [52]:

```
#Check whether prompt words are present
def is_question(words):
    for word in words:
        if word in['when','where','what','how','why','who']:
            return 1
        else:
            return 0
```

In [60]:

```
word_freq=gen_freq(dataset.clean_text.str)
#100 most rare words in the dataset
rare_100=word_freq[-100:]
#Number of words in a tweet
dataset['word_count']=dataset.clean_text.str.split().apply(lambda x:len(x))

#Negation present or not
dataset['any_neg']=dataset.clean_text.str.split().apply(lambda x:any_neg(x))

#Prompt present or not
dataset['is_question']=dataset.clean_text.str.split().apply(lambda x:is_question

#Any of the most 100 rare words present or not
dataset['any_rare']=dataset.clean_text.str.split().apply(lambda x:any_rare(x,rar

#Character count of the tweet
dataset['char_count']=dataset.clean_text.apply(lambda x:len(x))
dataset
```

Out[60]:

	id	label	tweet	clean_text	word_count	any_neg	is_question	any_rar
0	1	0	@user when a father is dysfunctional and is s...	@ ...	3	0.0	0.0	0.
1	2	0	@user @user thanks for #lyft credit i can't us...	@ @ # ...	6	0.0	0.0	0.
2	3	0	bihday your majesty		0	NaN	NaN	Na
3	4	0	#model i love u take with u all the time in ...	# ...	2	0.0	0.0	0.
4	5	0	factsguide: society now #motivation	: #	2	0.0	0.0	0.
...	...	...	...	...	...	...	...	
5237	31935	1	lady banned from kentucky mall. @user #jcpenn...	. @ # ...	4	0.0	0.0	0.
5238	31947	1	@user omfg i'm offended! i'm a mailbox and i'...	@ ! ...	5	0.0	0.0	0.
5239	31948	1	@user @user you don't have the balls to hashta...	@ @ ...	5	0.0	0.0	0.
5240	31949	1	makes you ask yourself, who am i? then am i a...	, ? ...	6	0.0	0.0	0.
5241	31961	1	@user #sikh #temple vandalised in in #calgary,...	@ # # # ,...	6	0.0	0.0	0.

5242 rows × 9 columns

In [ ]: