Name: Faris Chaudhry
Batch: LISUM25

Hate Speech Detection

Week 10

Team Member Details

Name: Faris Chaudhry

Email: faris.chaudhry@outlook.com

Country: United Kingdom

University: Imperial College London

Specialization: NLP

Problem Description

“The term hate speech is understood as any type of verbal, written or behavioural communication that attacks or uses derogatory or discriminatory language against a person or group based on what they are, in other words, based on their religion, ethnicity, nationality, race, colour, ancestry, sex or another identity factor.

Hate Speech Detection is generally a task of sentiment classification. So, for training, a model that can classify hate speech from a certain piece of text can be achieved by training it on a data that is generally used to classify sentiments. We will use the Twitter tweets to identify tweets containing Hate speech.”

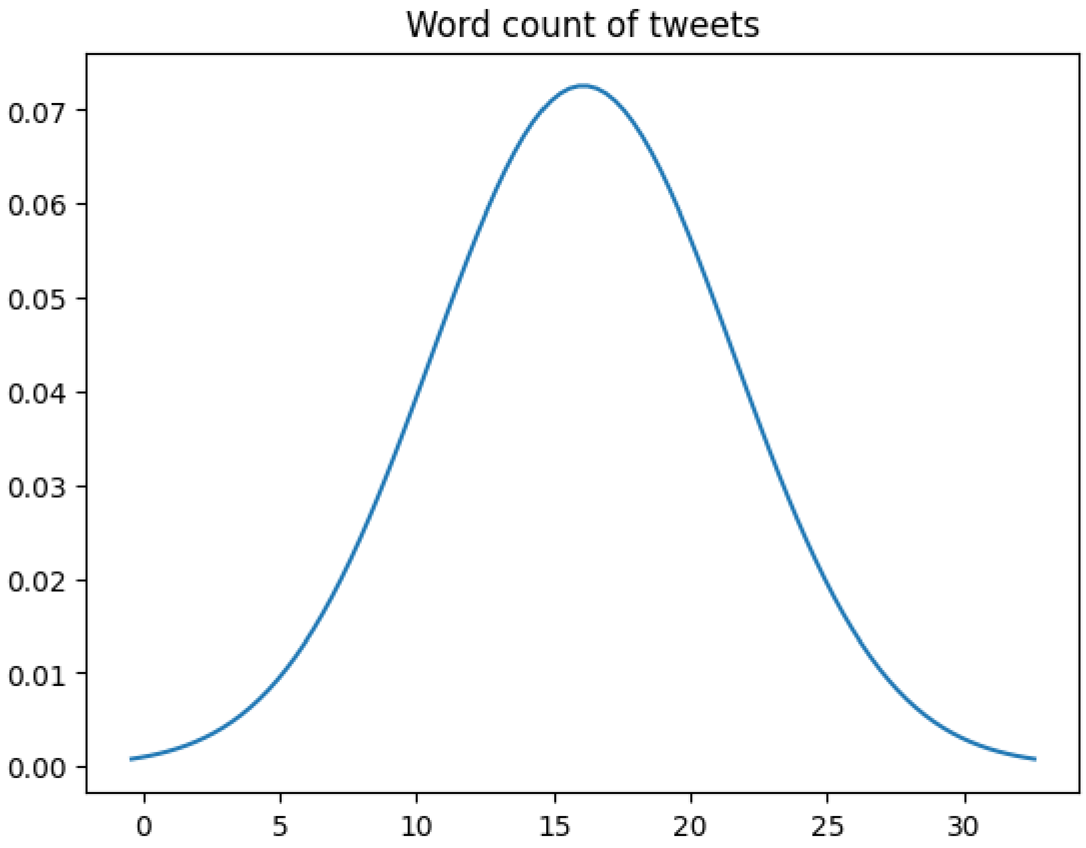
Name: Faris Chaudhry
Batch: LISUM25

Exploratory Data Analysis

Due to the nature of this task, there is not much possibility for EDA.

Word Count

Skewness is 0.15 which is quite low, so normal distribution is a suitable method.



Word Length (in Characters)

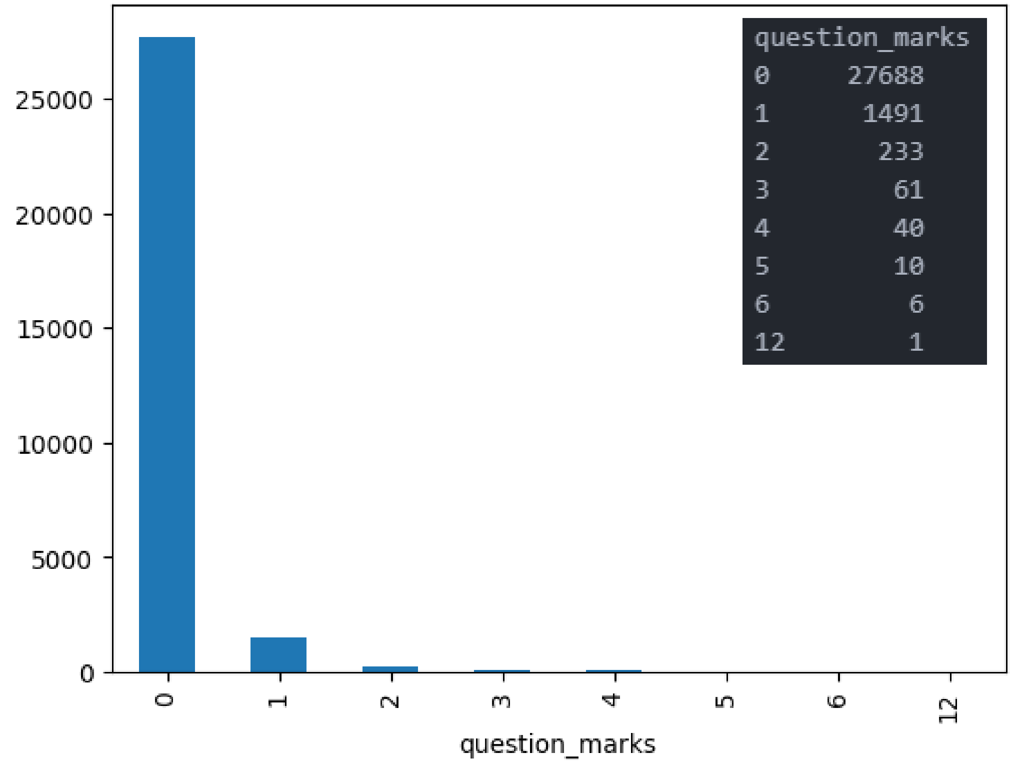
Mean: 5.60

Std. Dev.: 1.98

Skewness: 7

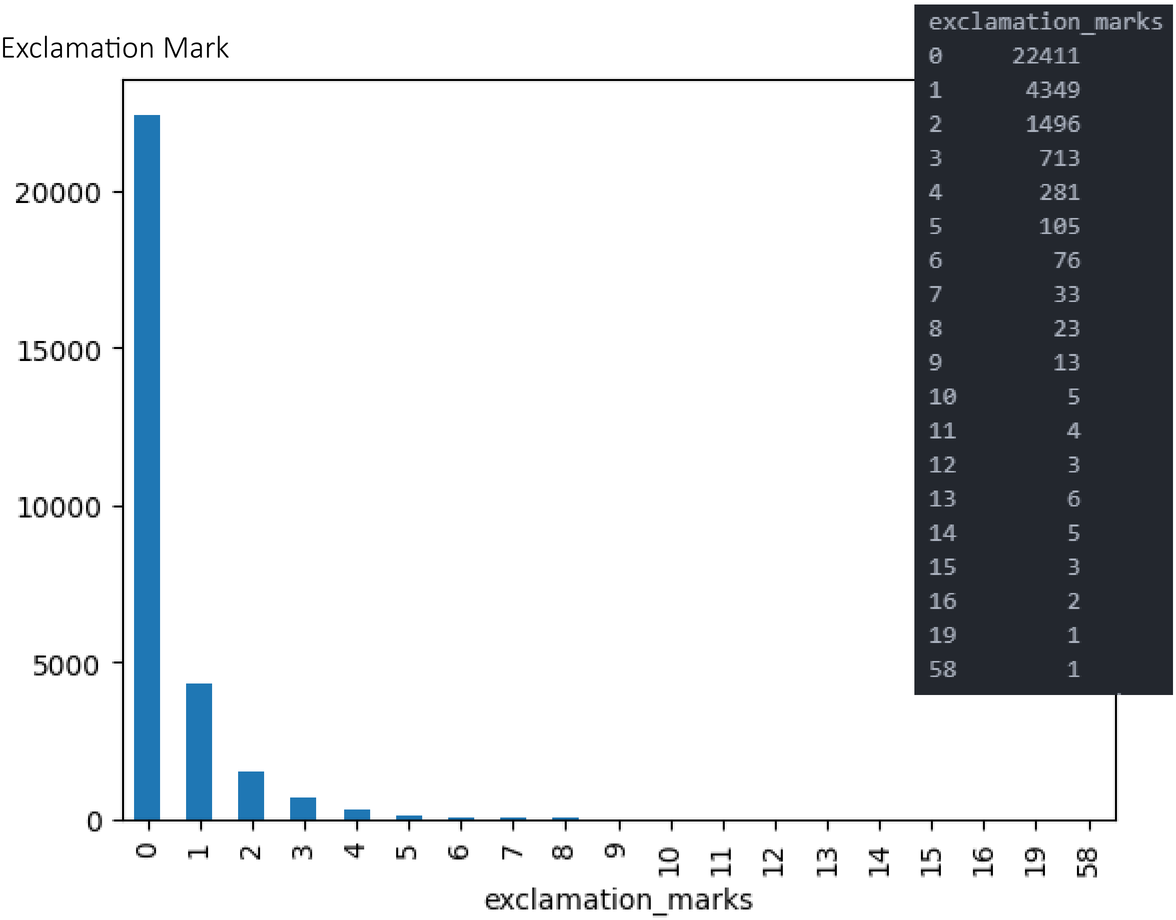
Name: Faris Chaudhry
Batch: LISUM25

Question Marks



```
question_marks
0      27688
1       1491
2        233
3         61
4         40
5         10
6          6
12         1
```

Exclamation Mark

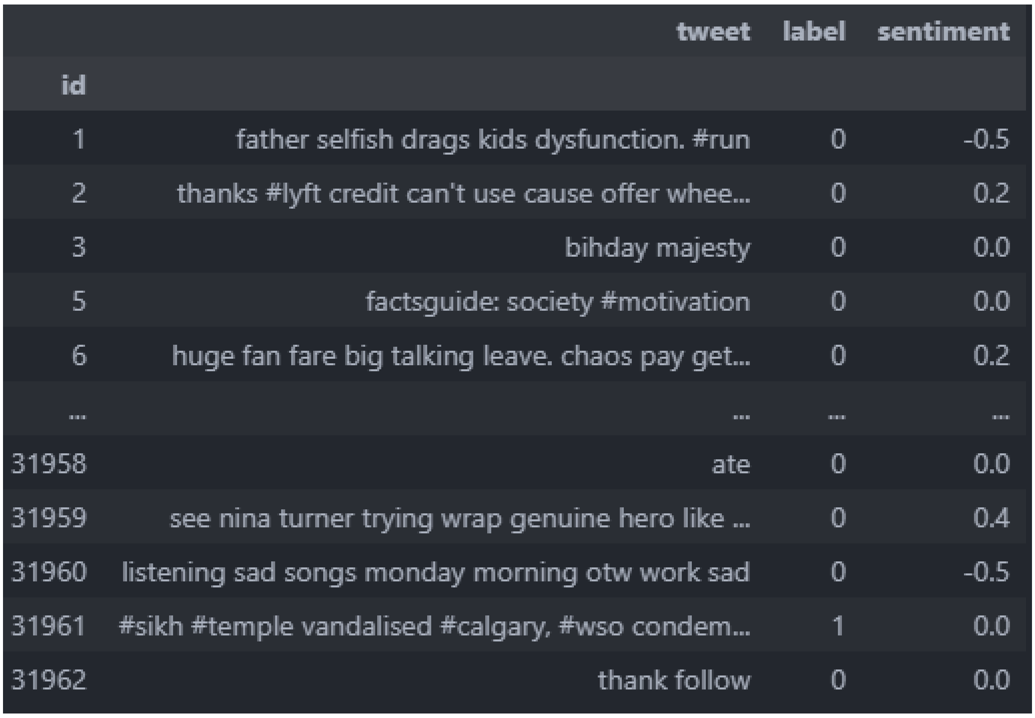


```
exclamation_marks
0     22411
1      4349
2      1496
3       713
4       281
5       105
6        76
7        33
8        23
9        13
10        5
11        4
12        3
13        6
14        5
15        3
16        2
19        1
58        1
```

Name: Faris Chaudhry
Batch: LISUM25

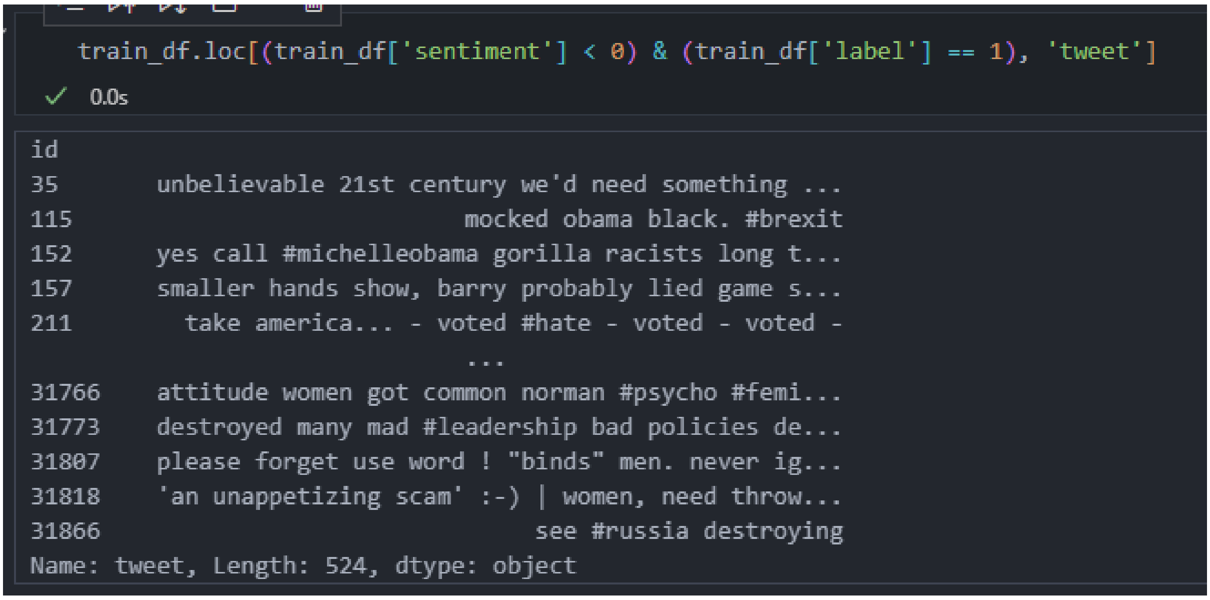
Sentiment

Using TextBlob we can assign a general sentiment to each tweets word list.



id	tweet	label	sentiment
1	father selfish drags kids dysfunction. #run	0	-0.5
2	thanks #lyft credit can't use cause offer whee...	0	0.2
3	bihday majesty	0	0.0
5	factsguide: society #motivation	0	0.0
6	huge fan fare big talking leave. chaos pay get...	0	0.2
...	...	...	...
31958	ate	0	0.0
31959	see nina turner trying wrap genuine hero like ...	0	0.4
31960	listening sad songs monday morning otw work sad	0	-0.5
31961	#sikh #temple vandalised #calgary, #wso condem...	1	0.0
31962	thank follow	0	0.0

By filtering for negative sentiments with hate speech labels, we can see that there are 524 such entries (approximately 25% of labelled data). However, there are many false positives which means the sentiment is more of criticism in general than hate speech specifically. Overall, the correlation between sentiment and the label is -0.14.



```
train_df.loc[(train_df['sentiment'] < 0) & (train_df['label'] == 1), 'tweet']
```

✓ 0.0s

```
id
35       unbelievable 21st century we'd need something ...
115                              mocked obama black. #brexit
152      yes call #michelleobama gorilla racists long t...
157      smaller hands show, barry probably lied game s...
211        take america... - voted #hate - voted - voted -
                               ...
31766    attitude women got common norman #psycho #femi...
31773    destroyed many mad #leadership bad policies de...
31807    please forget use word ! "binds" men. never ig...
31818    'an unappetizing scam' :-) | women, need throw...
31866                               see #russia destroying
Name: tweet, Length: 524, dtype: object
```

TF-IDF

- Term frequency (TF): Measurement of how frequently a term occurs within a document. A word with occurrences in multiple tweets should have more significance.

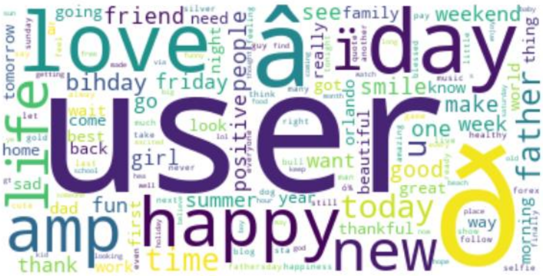
Name: Faris Chaudhry
Batch: LISUM25

- Inverse document frequency: IDF (word) = log(number of entries / number of entries containing word). A word appearing too often isn't very significant because it only adds statistical noise.
- TF-IDF: Reduces significance of commonly occurring words even if they have a high idf.

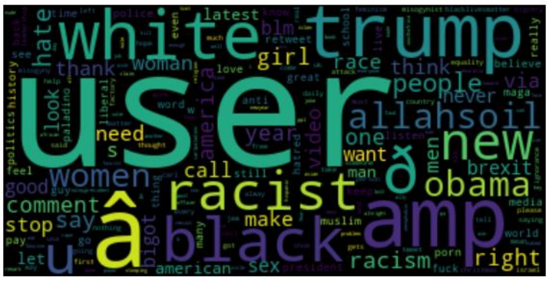
However, since the words are already tokenized, tf and idf have a very high correlation, so keeping both only increases dimensionality.

TextCloud of common words

Non-Hate Comments



Hate Comments


user
white
trump
amp
black
racist
â
new
allahsoil
obama
women
people
hate
girl
race
comment
stop
racism
right
brexit
america
american
sex
call