Hate Speech Detection

Week 10

# Team Member Details

Name: Faris Chaudhry

Email: [faris.chaudhry@outlook.com](mailto: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.”

# 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.

A graph of a number of tweets

Description automatically generated

### Word Length (in Characters)

Mean: 5.60

Std. Dev.: 1.98

Skewness: 7

### Question Marks

A screenshot of a black background with white numbers

Description automatically generatedA graph with blue bars

Description automatically generated

A screenshot of a black background with numbers

Description automatically generated

### Exclamation Mark

A graph of a number

Description automatically generated

### Sentiment

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

A screenshot of a social media message

Description automatically generated  
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.

A screenshot of a computer screen

Description automatically generated

### 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.
* 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.