Week 10

EDA Analysis

Team Members:

Member no.	Name	Email	Country	college	Specialization
1	Zyad Hussein	Zyadrashad262@yahoo.com	Egyptian	University of York	NLP

Problem description:

Hate speech across social media platforms has been significantly increasing leading to the generation of insecurity, increasing unsafeness feelings, and limiting the freedom of expression. Hate speech can be expressed in many forms such as tweets, words, sentences, paragraphs and many other forms of expression. Hate speech can target users with different ethnicities, race, sex, backgrounds, nationalities, colour, or ancestry which can produce adverse long-term effects on the targeted group. Hereby, it is believed that a solution is needed to address the issue. Such task isn't possible to tackle manually. Therefore, it is believed that an implementation of a deep learning model that automatically recognises and removes hate speech without the interfering of any humans can significantly reduce the proposed issue and overcome the issue. The elimination of hate speech can help increase the contribution of more users to topics and allow them to freely discuss any topics with no fear of hate attacks. Such implementation could allow new age group users to join social media platforms and experience it. A data was collected to train the deep learning model to autonomously detect hate speech when a speech is presented to it. EDA analysis was performed on this dataset to identify the percentage of hate speech in twitter tweets and the most frequent words used in hate speech to identify types of hate speech used.

EDA Analysis:

1. It was first decided to identify the average tweet length in twitter tweets to reveal more characteristics of the tweets structure. After performing analysis, it was found that average tweet lies within 85-89 words as shown in figure 1. Positive speech was found to be utilizing less words than negative speech as shown in figure 2.

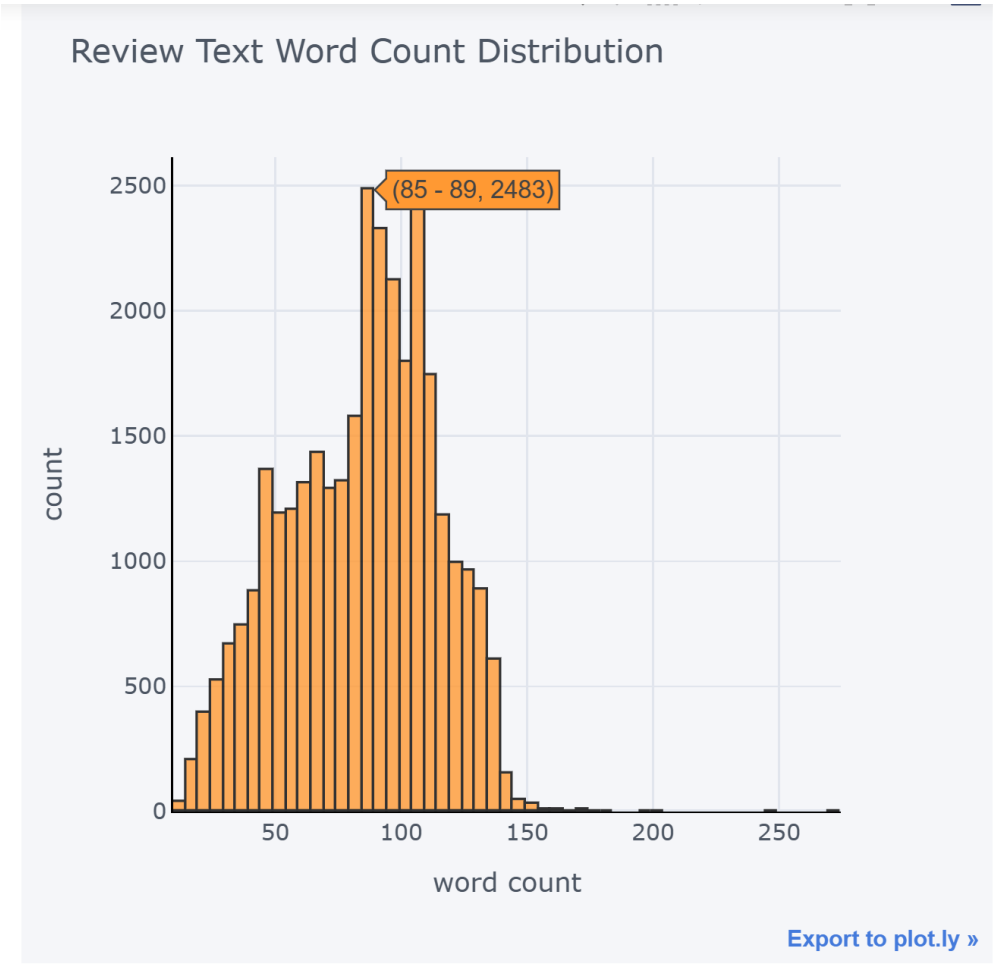

Fig 1: frequency of highest word/character count.

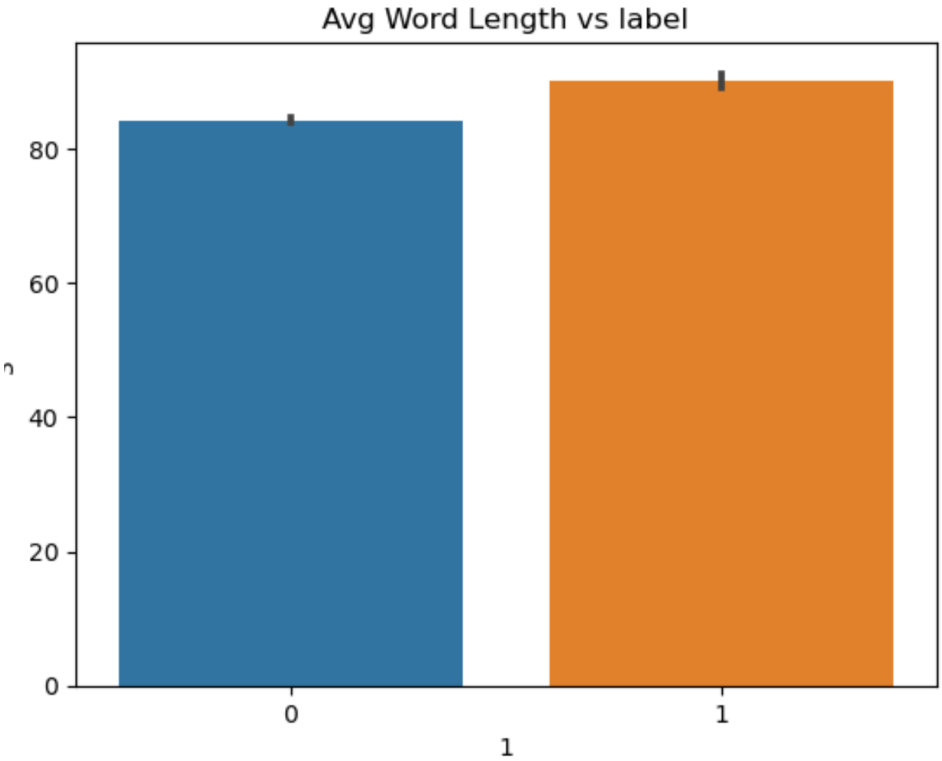

Fig 2: Average word count in positive and negative speech.

2. It was decided to vectorize all the words in negative hate speech to identify the type of hate words used in the negative speech and identify the targeted group of users. Figure 3 provides the amount of words used and the frequency of each word used in the negative speech part of dataset. according to the presented plot, it is believed that a significant amount of hate speech is directed towards colour such as 'white ' and 'black' and towards specific people with their names such as 'Trump' and 'Obama'. The word racist has also been noticed more frequently in the negative hate speech which is a significant evidence of hate speech presented.

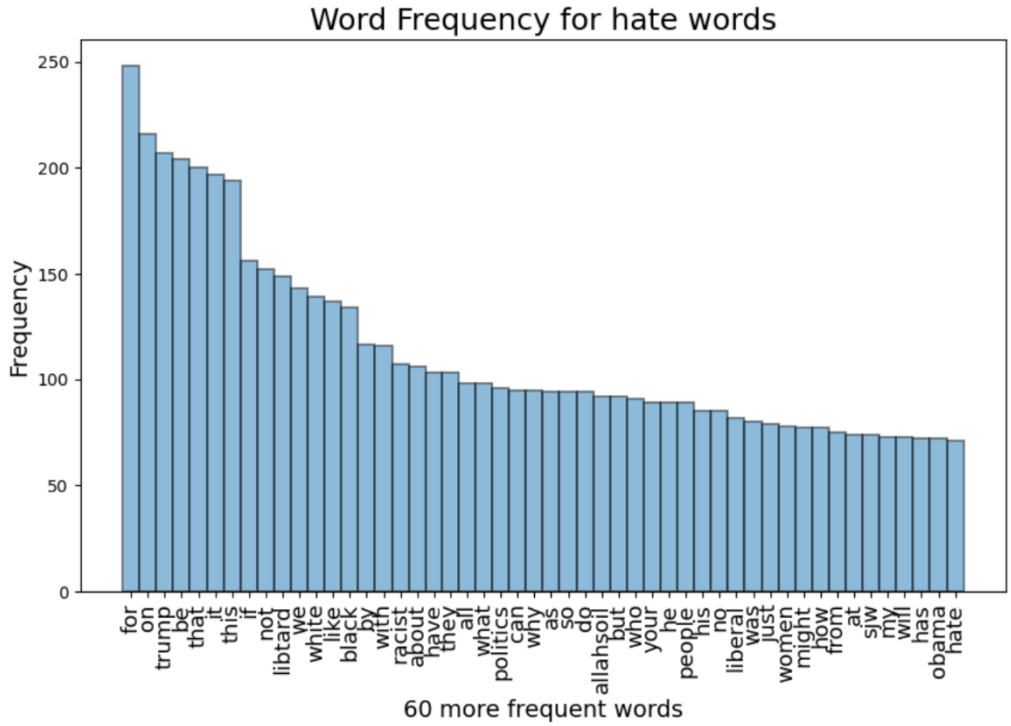

Fig 3: frequency of hate speech words used.

3. The number of negative tweets was also identified through a plot and calculations of labels '1' which represents the negative labels. The number of negative speeches was lower than positive speeches. However, the ratio of some of the negative words used within the hate tweets was significantly high, meaning that a specific audience is targeted more frequently than others. Presentation will also provide more interpretation to all these mentioned parts.

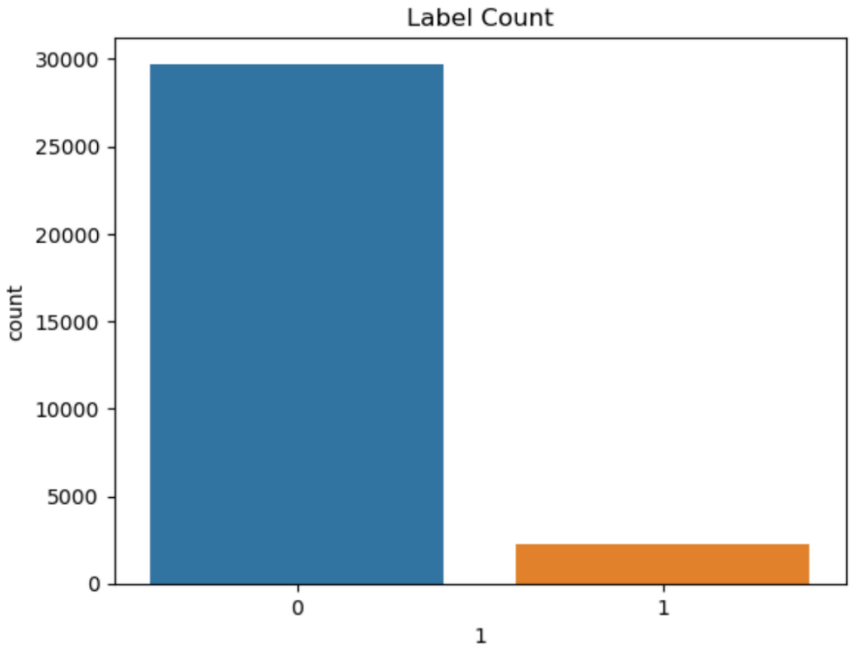

Fig 4: negative speech count and positive speech count.

Final Recommendation:

It is believed that NLP utilization can aid in removing the hate speech discovered as the percentage of hate speech has been increasing and it is targeting specific users which could dramatically result in adverse effects. The illustrated data revealed that a specific group f colour is being targeted which requires a solution to such issue to protect that group and any other minor groups from ate speech attacks.