Twitter Sentiment Analysis

Milestone 1 Report

Problem

The aim of this project is to extract the sentiment of tweets and classifying it as either a positive or a negative sentiment. A solution to this problem is important to a wide variety of domains as will be discussed in the following section.

Client

Sentiment analysis is a multidisciplinary necessity. The following are only some of the potential uses of sentiment analysis.

- **Online Commerce:** The sentiments can be extracted from the reviews that the people give about products to understand the general like or dislike of a certain feature or a product. This can be used to make improvements to the particular products
- **Voice of the Market (VOM):** Voice of the Market is about determining what customers are feeling about products or services of competitors. Accurate and timely information from the Voice of the Market helps in gaining competitive advantage and new product development. Detection of such information as early as possible helps in direct and target key marketing campaigns
- **Voice of the Customer (VOC):** Voice of the Customer is concern about what individual customers are saying about products or services. It means analyzing the reviews and feedback of the customers. VOC is a key element of Customer Experience Management. VOC helps in identifying new opportunities for product inventions. Extracting customer opinions also helps identify functional requirements of the products and some non-functional requirements like performance and cost.
- **Brand Reputation Management:** Brand Reputation Management is concern about managing your reputation in market. Opinions from customers or any other parties can damage or enhance your reputation. Brand Reputation Management (BRM) is a product and company focused rather than customer. Now, one-to-many conversations are taking place online at a high rate. That creates opportunities for organizations to manage and strengthen brand reputation.
- **Government:** Sentiment analysis helps government in assessing their strength and weaknesses by analyzing opinions from public

Dataset

The data that I will be using is of 100,000 tweets each marked with either a positive or a negative tweet. The dataset can be easily downloaded from here.

The dataset is available as a csv file and has essentially 2 columns of interest.

1. Sentiment - 0,1 (Negative,Positive)
2. Tweet - The text associated with the sentiment

Data Preprocessing

The tweets were processed according to the following and in the same order.

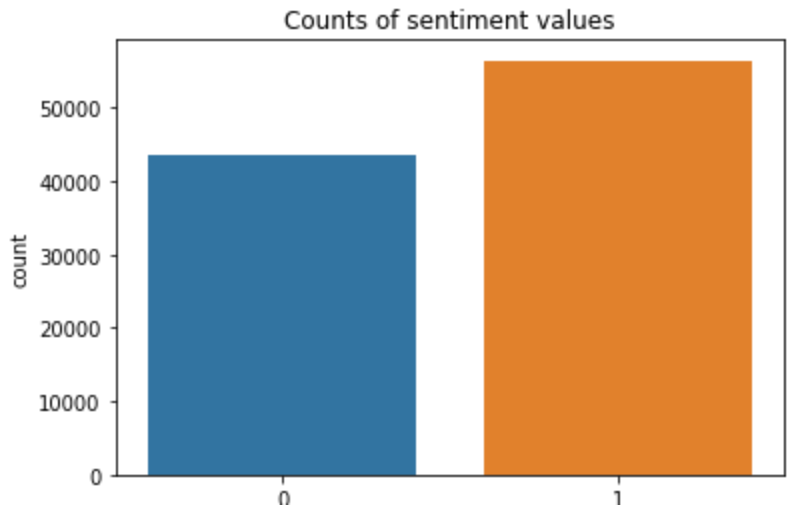
1. Tags removal (i.e '@')
2. Lowercasing
3. Numbers removal
4. HTTP links removal
5. Emojis were processed. Emojis were labelled EMO_POS or EMO_NEG
6. Punctuation removal
7. Removed extra white spaces
8. Words which did not consist of alphabets were removed
9. Stop words were removed (These words add no value to the sentiment of the tweet e.g The, He etc)
10. Character repetitions were removed e.g *funnny* was changed to *funny*
11. Words were lemmatized to bring to their basic form e.g adventurous changed to adventure

After performing these steps I checked to see if there were any tweets which were reduced to an empty string. I could not find any such cases.

Exploratory Data Analysis

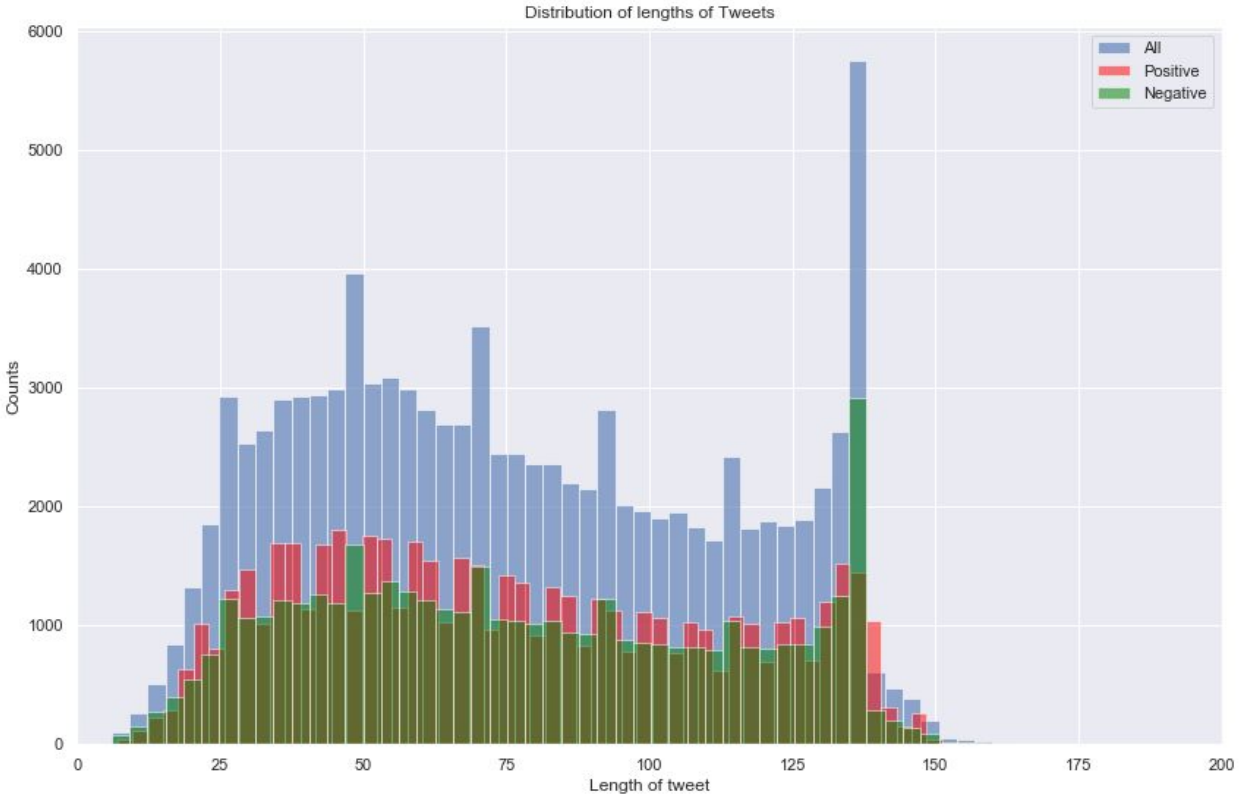
Distribution of sentiments

As can be seen below, both sentiments are somewhat balanced



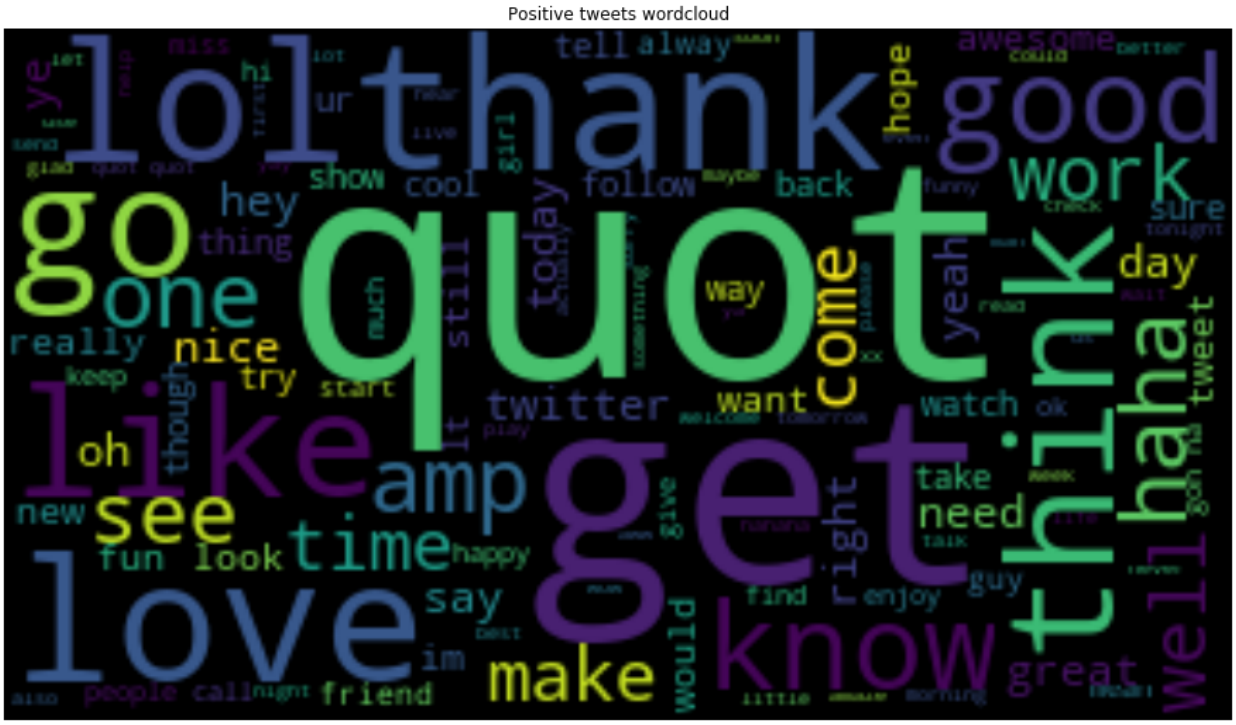
Distribution of lengths of tweets

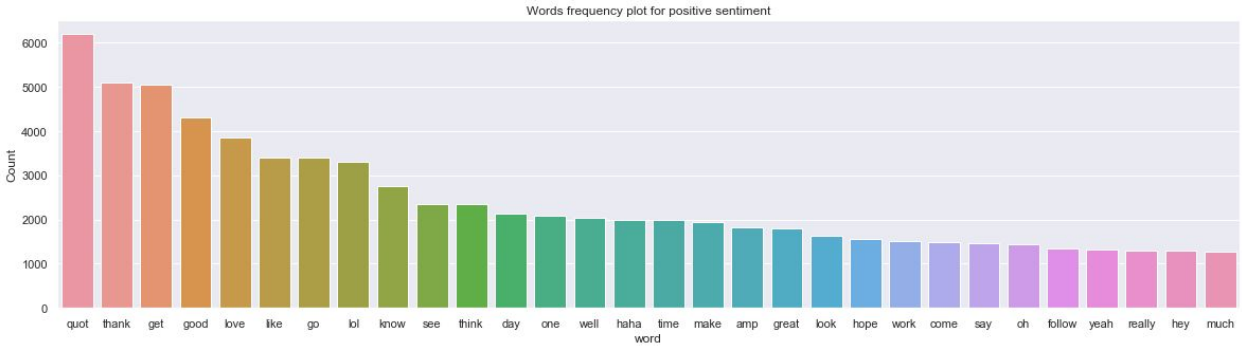
We can also see below that the lengths of both the sentiments follow the same distribution.



Word Clouds and Word frequency plots

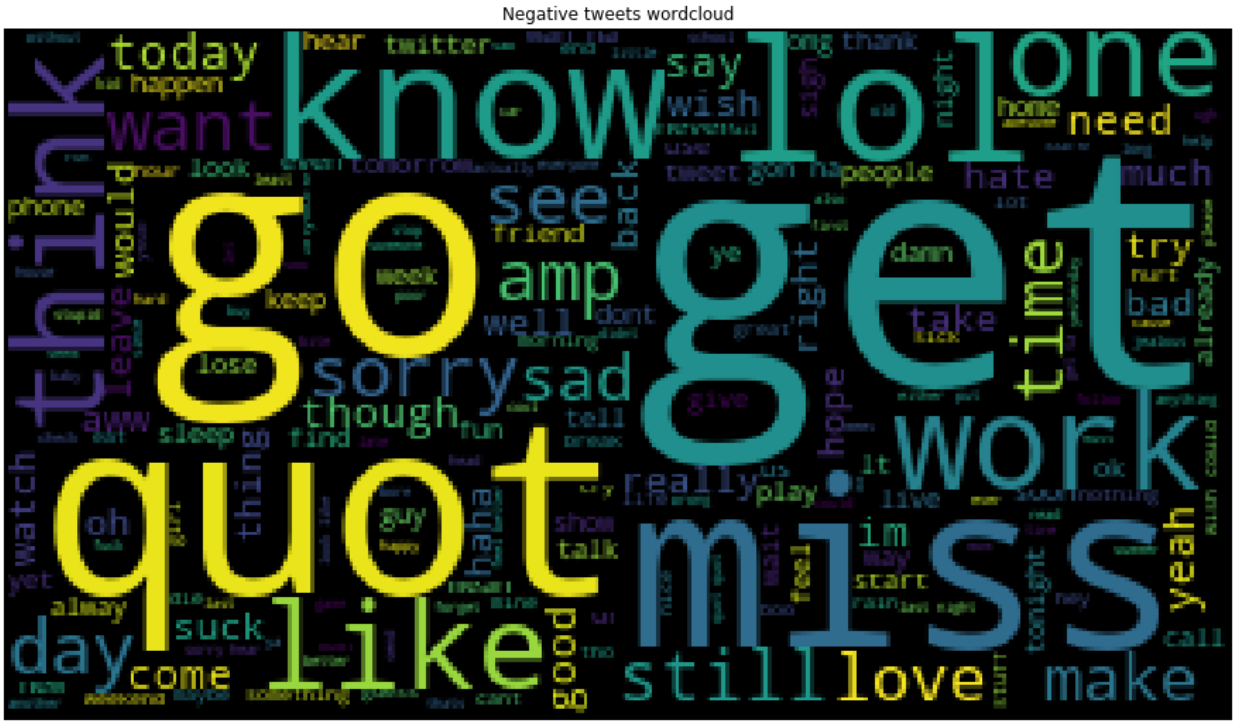
Positive Sentiments

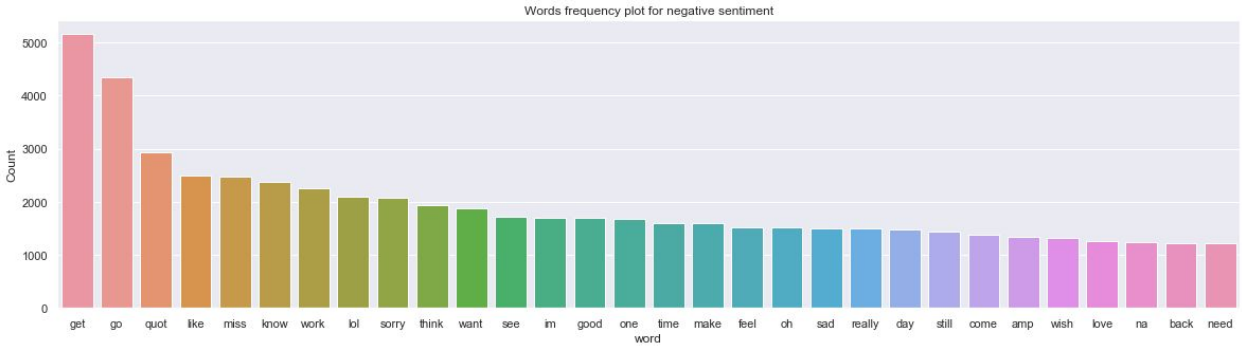




We can see that for the positive sentiment words like lol,thank, love, good show up more frequently.

Negative Sentiments





For the negative sentiment words like miss, work, sad, wish are more frequent.