Capstone 2- Milestone Report : 05/06/2020
Prediction of Twitter gender based on Tweet text

Objective and Goal:

Sentiment analysis is famous among major brands. It is the way to identify the tone and emotions expressed through written or spoken online communication. It is famous with big companies like

- Twitter
- Amazon
- Facebook
- Netflix.

Based on the dataset we have obtained, there was some analysis which was done based out of the following questions.

- ➢ How well do words in tweets and profiles predict user gender?
- ➢ What are the words that strongly predict male or female gender?
- ➢ How well do stylistic factors (like link color and sidebar color) predict user gender?

The goal of this project is to simply view a Twitter profile and judge whether the user was a male, a female, or a brand (non-individual). NLP is used and its different methods paves way for achieving a solution for performing the analysis. This is useful for the prospective client in determining a particular gender in analyzing well, based on seeing a user's tweet or a profile.

Dataset:

This data set was used to train a CrowdFlower AI gender predictor. The dataset contains 20,000 rows, each with a username, a random tweet, account profile and image, location, and even link and sidebar color.

Twitter User Gender Classification

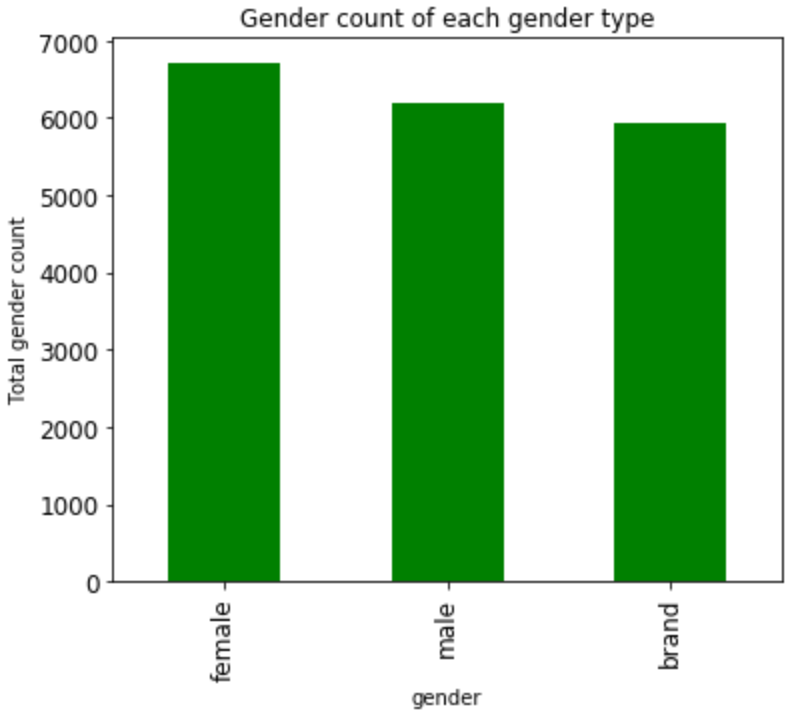
Data Wrangling:

In this part, there have been some NaN's in the dataset. The text and description columns contain loads of information regarding a person's tweet and also, there are some other columns like retweet counts, sidebar link color etc., which might be useful in determining the gender.

Also, there are columns which were not much useful to us in gender prediction. Features like **profile_yn, profile_yn_gold, golden, trusted_judgements** were dropped from the sub DataFrame.

Data Story:

I have calculated the different gender type counts from the dataset.



The above bar plot shows the gender count type of how many genders are present in the dataset. There is an unknown gender located as well in the dataset, which was not needed. So, I have removed those values, and retained only the useful information.

Besides this, the tweet counts for each gender was also calculated. This was useful in knowing which gender has more tweet counts.

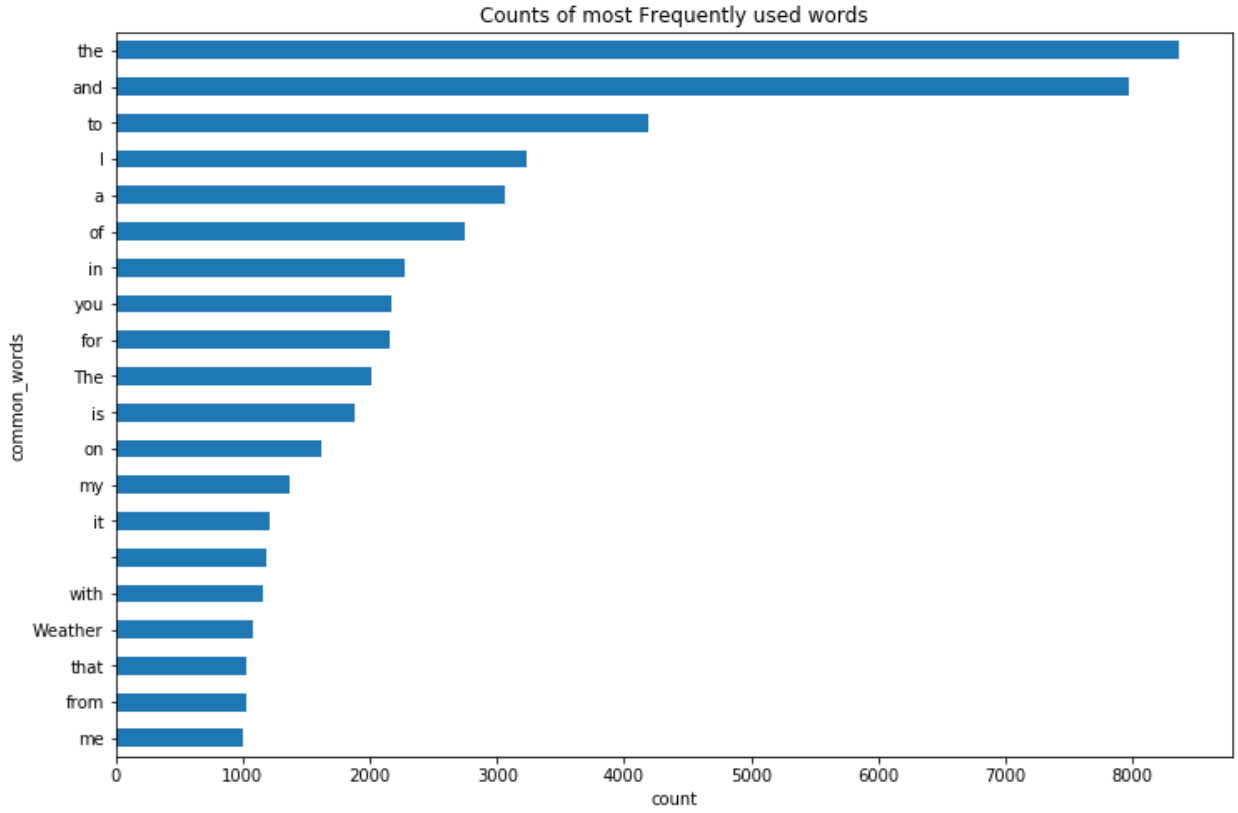
Total female tweets: 5725

Total male tweets: 5469

Total brand tweets: 4328

Data Visualization/ Manipulation of Text Data:

We now need to somehow manipulate the text data for our data to preprocess and be useful for our models later. NLP contains a library called NLTK. It stands for Natural Language Toolkit. This toolkit is one of the most powerful NLP libraries which contains packages to make machines understand human language and reply to it with an appropriate response. Tokenization, Stemming, Lemmatization, Punctuation, Character count, word count are some of these packages present.



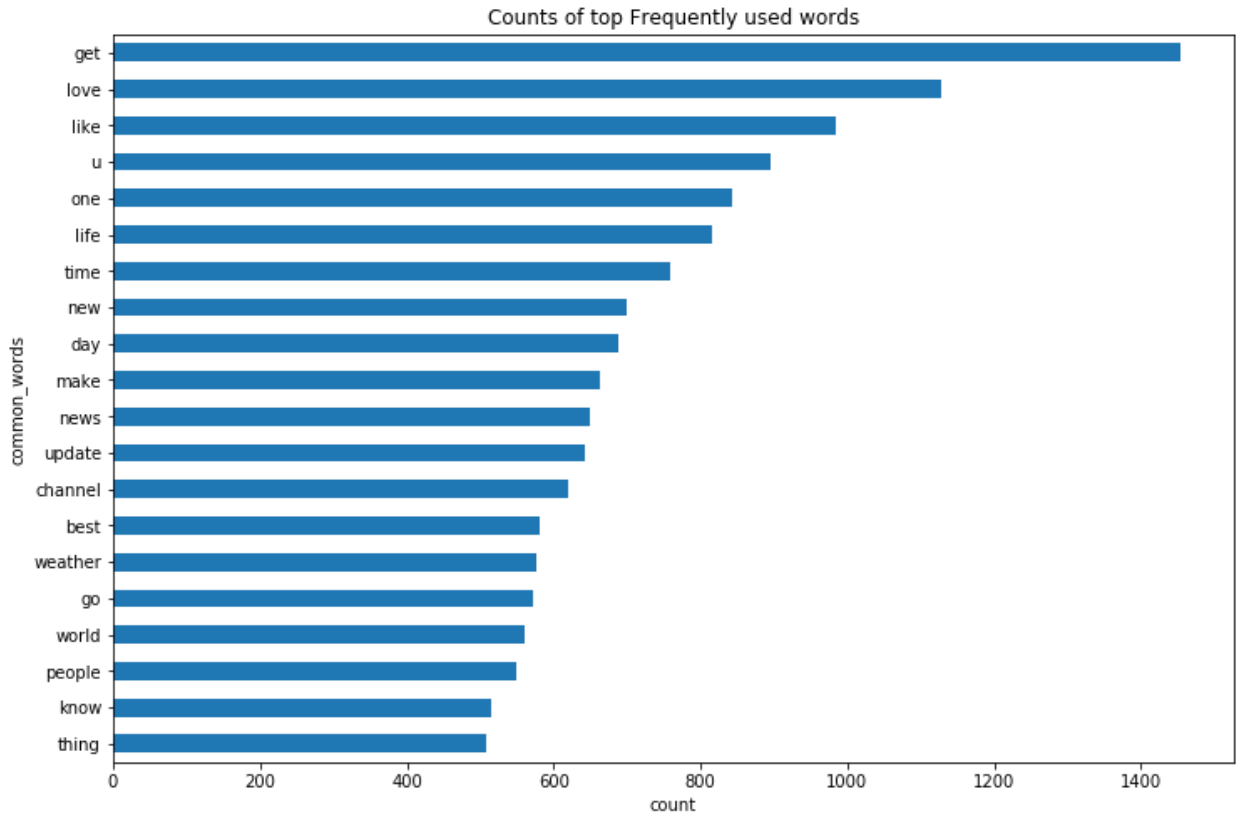
The plot shows the most common words count frequently being used throughout the dataset text feature. Later on, the text and description features get combined to impute for the missing values. Also, we might get rich content of data from the combined features. But the above plot contains only normal English language words like a, the, an etc., which is not much from the ‘text’ feature. We will shortly introduce the stopwords concept that can reduce this drawback.

Removing stop words and applying Lemmatization:

A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore. In natural language processing, useless words (data), are referred to as stop words.

Lemmatization is also stemming but produces results which are all valid words.

Also, a regular expression is used below to specify a set of strings that matches it; they allow you to check if a particular string matches a given regular expression.

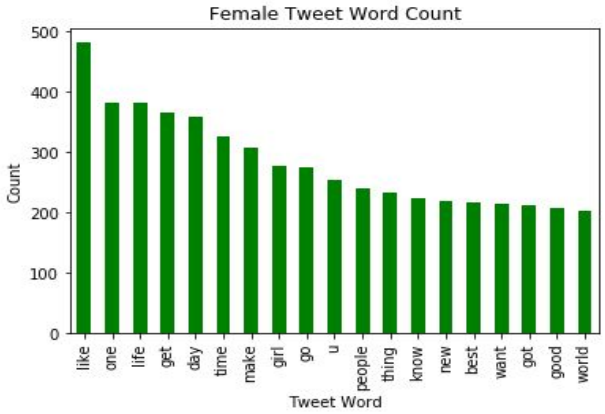


The plot now is cleaned up using regular expressions, tokenizing and using stopwords to the combined features of text and description. We will next see how to generate the most common frequency of words using concept called **'WordCloud'.**

Wordcloud:

Ususally, whenever we deal with NLP or text, we see a cloud filled with lots of words in different sizes, which represent the frequency or the importance of each word. Wordcloud is the exact representation of this. It is a data Visualization technique used for text analysis.

From the image, we now see that the words new, love, life, one etc.. are displayed in a large font, meaning that these words occur very frequently and the frequency for these words is more.



The bar plot shows the female Tweet word count. The frequency is described as to which word occurs more frequently among female gender. The same applies to male gender as well on how frequent the word is common and how many times it appears.