**Twitter Sentiment Analysis:**

**Problem Statement:** Need to predict the hate speech in the tweets.

Approach:

1. Clean the data and Analyse it
2. Build the models
3. Compare the models and select the best one
4. Based on the best model, predict the “Test” data set

The data is in the form of text. This type of data cannot be used for model building. It needs to be converted into number format to employ the algorithms on the model. But, before we employ any algorithm, we need to clean the data as it contains many errors wrt to grammar, tenses, special characters etc.

In NLP, we tackle this problem by different methods.   
  
Lemmatization -> Removes the inflectional endings of words

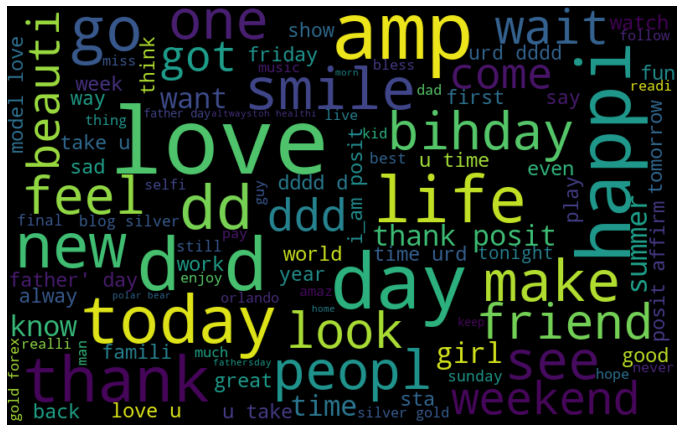
Stemming -> Reduces the inflected word

Stop Words -> Need to remove them

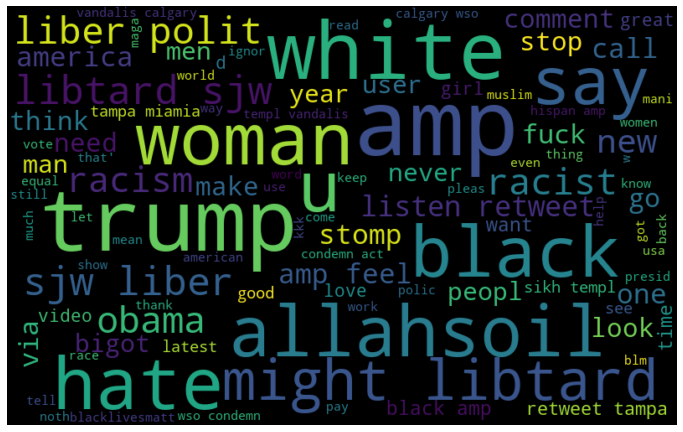
**Positive Words and Negative Words:**

By going for a visualisation, we get an intuitive feel about the type of words that are mostly used in Positive tweets and Negative tweets.

Positive Ones:



Negative Ones:



Now, we have fair understanding about the Positive and Negative Tweets. As mentioned earlier, to build a model, we can’t build directly on words. First we need to convert them into Numeric form.

Here, we do it by TF-IDF.

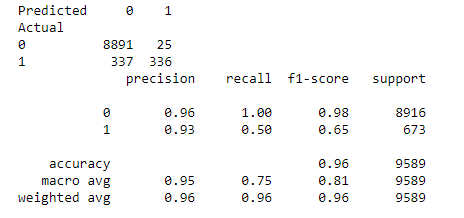
**What is TF-IDF?**

TF-IDF is an information retrieval technique that weighs a term’s frequency (TF) and its inverse document frequency (IDF). Each word or term has its respective TF and IDF score. The product of the TF and IDF scores of a term is called the TF\*IDF weight of that term.

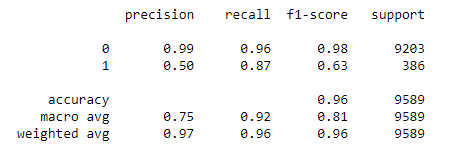
Once after converting them into TF-IDF form, we can employ our models on that. To get to the best model, the data set is divided into two parts, train and test (randomly). The model which we get out of train data, is deployed on the test set to find out the model accuracy.   
  
**Different models applied and the result of the models:**

I modelled the training data using Random Forest and Logistic Regression. The below are the values I got after validating the model with test set.

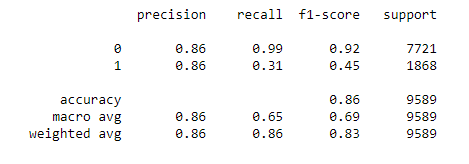
**#Random Forest**

****

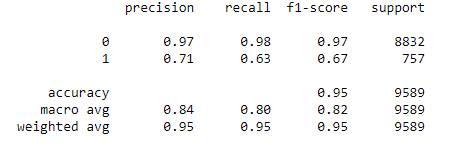
**# Logistic Regression**

****

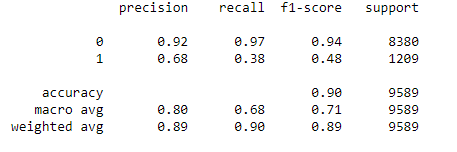
**#Random under-sampling with imblearn**

****

**#Under-sampling: Tomek links**

****

**#Synthetic Minority Oversampling Technique (SMOTE)**

****

The Best model I could observe is Random Forest model. Eventhough the recall value is low, the F1-score and negative comments prediction is higher than the other models. So, I am predicting the test set with Random forest model.