Name: Geeth Chandra Kalleda

Two Level Multi Label Classification

Background:

With the incremental use of emails as an essential and popular communication tool over the Internet, there comes a serious threat that impacts the Internet and the society. This problem is known as spam.
The aim of this project is to build a spam detector in the first level and classify an email as Spam/Not Spam. The second level of classification is to do a semantic analysis on these Spam and Not Spam messages.
Finally after this classification, we have 2 x 3 labels - one on spam status and another on sentiment. A correlation would be analysed between the spam and sentiment.

Problem:

The major challenge was to increase the accuracy by vigorous text processing and also combine two levels of classification on the same dataset.

Required Tools & Libraries:

Numpy, pandas, sklearn, nltk, TextBlob, matplotlib
Logistic Regression model is used for Spam detection.

Input :

A dataset of 5572 records with 2 columns. One column is the email message and the other column is labelled against spam/ham.

Output:

A correlation is established which tries to explain the following questions:

- Are spams mostly positive ?
- Can spams be predicted with high accuracy from sentiments ?
- Are standard emails generally positive?

World cloud is used to represent all the 2 x 3 labels.

Approach:

Input data is classified on two levels - One on Spam/ham and other on Sentiment.

First level of classification on Spam/ham:

Vigorous text processing is done on input data which includes removing stop words, stemming the words and normalizing the length of text.
A logistic regression model is trained using the train data and then later used to predict on test data.

Second level of classification on Sentiment:

Once an email is classified as spam/ham, it is now predicted if it is positive/negative/neutral using a pretrained sentiment analyser (TextBlob).

Once classification on two levels are done, we get a total of 6 labels:
Spam positive, Spam negative, Spam neutral, Ham positive, Ham negative, Ham neutral.

Now we can analyse the relation between Spam/ham and sentiment.

Evaluation:

The total test messages on which I have tested are 1672 messages.

Total number of spam messages among test data – 184

Total number of ham messages among test data - 1488

Spam Positive - 104

Spam Negative - 15

Spam Neutral - 65

Percentage of spam positives among spam - 56.52%

Percentage of spam negatives among spam - 8.15%

Percentage of spam neutrals among spam - 35.32%

Ham Positive - 600

Ham Negative - 255

Ham Neutral - 633

Percentage of Ham positives among Ham - 40.32%

Percentage of Ham negatives among Ham - 17.13%

Percentage of Ham neutrals among Ham - 42.54%

Discussion:

1. 56.52% of spam messages are positive and 35.52% of spam messages are neutral. Only 8% of spam are negative.
   So we can derive from the results that spam messages are mostly positive or neutral. Positive emails are greater than neutral emails among spams.

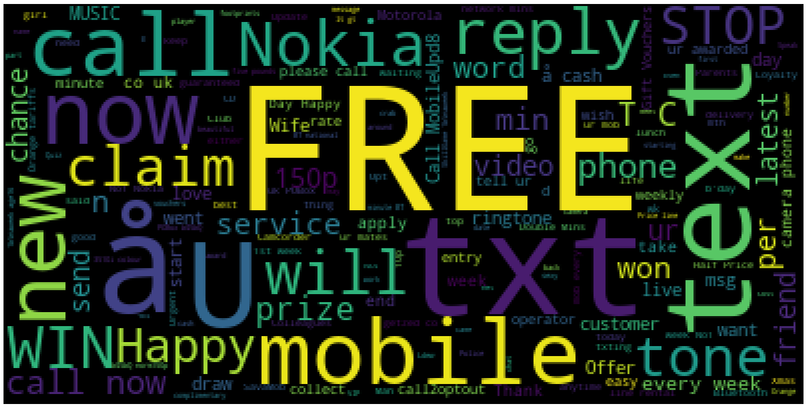
2. There are a total of 704 positive messages of which 104 are spam and the rest are ham.
   There are a total of 270 negative messages of which 15 are spam and 255 are ham.
   So, it is not possible to derive spam/ham from sentiment.

3. Of the Ham messages, most of them are positive or neutral and to be more accurate, neutral messages are slightly greater than positive messages.
   This indicates that the standard emails are generally neutral or positive.

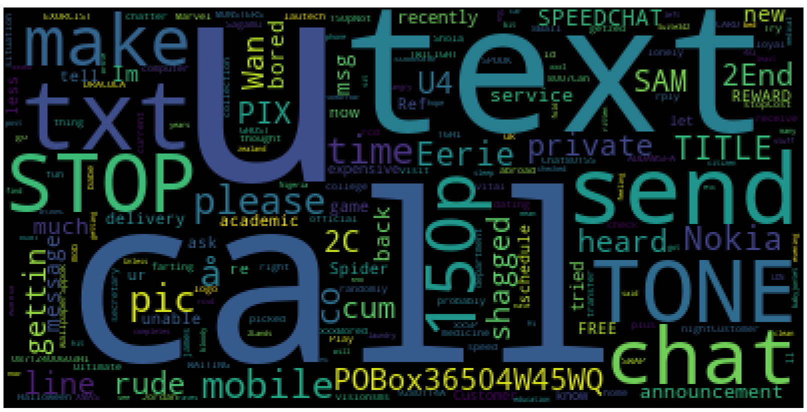
All the above results were based on a test data of 1672 messages.

Word maps for all the 6 labels are as follows:

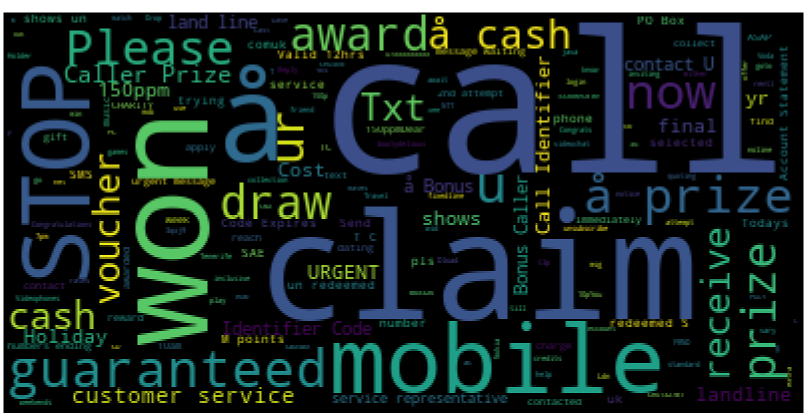
1. Spam positive:



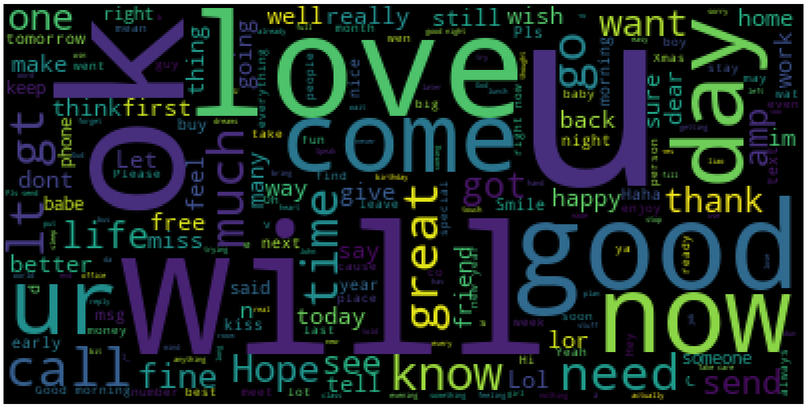
2. Spam Negative:



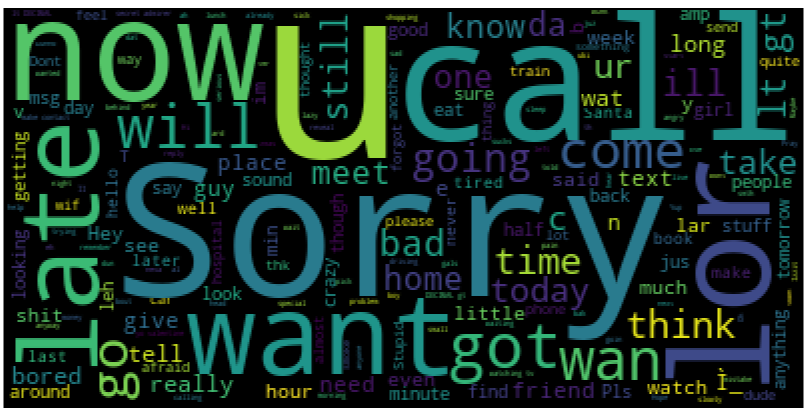
3. Spam Neutral:



4. Ham Positive



5. Ham Negative



6. Ham Neutral: