SPAM MESSAGE DETECTOR

Avadhoot V Pawaskar

RU ID: 197005413

Course No: 22:548:688

MITA Capstone Project

Under the guidance of

Professor. Michail Xyntarakis

Table of Contents

1. ABSTRACT

Over recent years, as the popularity of mobile phone devices has increased, Short Message Service (SMS) has grown into a multi-billion dollars industry. E-mail spam is a very recent problem for every individual. The e-mail spam is nothing; it is an advertisement of any company/product or any kind of virus which is received by the email client mailbox without any notification. To solve this problem the different spam detection techniques are used. In this project, we are using the Naïve Bayesian Classifier for spam classification and I will be using Scikit-learn, a popular machine learning library, to achieve our goals. Finally, I built a flask app and deployed it on Heroku. The Multinomial Naïve Bayesian Classifier is a very simple and efficient method for spam classification. Here we are using the spam dataset for classification of spam and non-spam mails. The feature extraction technique is used to extract the feature. The result is to increase the accuracy of the system.

Keywords

Spam Detection, SMS, Classification, Content Features

2. Introduction

In recent years, the internet has become an integral part of life. With increased use of the internet, numbers of email users are increasing day by day. This increasing use of email has created problems caused by unsolicited bulk email messages commonly referred to as Spam. Email has now become one of the best ways for advertisements due to which spam emails are generated. Spam emails are the emails that the receiver does not wish to receive. Many identical messages are sent to several recipients of email. Spam usually arises because of giving out our email address on an unauthorized or unscrupulous website .There are many of the effects of Spam. Fills our Inbox with a number of ridiculous emails. Degrades our Internet speed to a great extent. Steal useful information like our details on your Contact list. Alters your search results on any computer program. Spam is a huge waste of everybody's time and can quickly become very frustrating if you receive large amounts of it. Identifying these spammers and the spam content is a laborious task. Even though an extensive number of studies have been done, yet so far, the methods set forth still scarcely distinguish spam surveys, and none of them demonstrate the benefits of each removed element compose. Despite increasing network communication and wasting a lot of memory space ,spam messages are also used for some attacks . Spam emails, also known as non-self, are unsolicited commercial or malicious emails, sent to affect either a single individual or a corporation or a bunch of people. Besides advertising, these may contain links to phishing or malware hosting websites found out to steal confidential information.

Email spam targets singular clients with regular postal mail messages. Email spam records are regularly made by checking Usenet postings, taking Internet mailing records, or scanning the Web for locations. Email spams normally cost clients cash out-of-pocket to get. Numerous individuals - anybody with measured telephone administration - read or get their mail while the meter is running, as it were. Spam costs them extra cash. On top of that, it costs cash for ISPs and online administrations to transmit spam, and these expenses are transmitted specifically to endorsers.

3. DATASET

The files contain one message per line. Each line is composed of two columns: v1 contains the label (ham or spam) and v2 contains the raw text. This corpus has been collected from free or free for research sources on the Internet.

a) A collection of 425 SMS spam messages was manually extracted from the Grumble text Web site. This is a UK forum in which cell phone users make public claims about SMS spam messages, most of them without reporting the very spam message received. The identification of the text of spam messages in the claims is a very hard and time-consuming task, and it involved carefully scanning hundreds of web pages.

b) A subset of 3,375 SMS randomly chosen ham messages of the NUS SMS Corpus (NSC), which is a dataset of about 10,000 legitimate messages collected for research at the Department of Computer Science at the National University of Singapore. The messages largely originate from Singaporeans and mostly from students attending the University. These messages were collected from volunteers who were made aware that their contributions were going to be made publicly available.

c) A list of 450 SMS ham messages collected from Caroline Tag's PhD Thesis available.

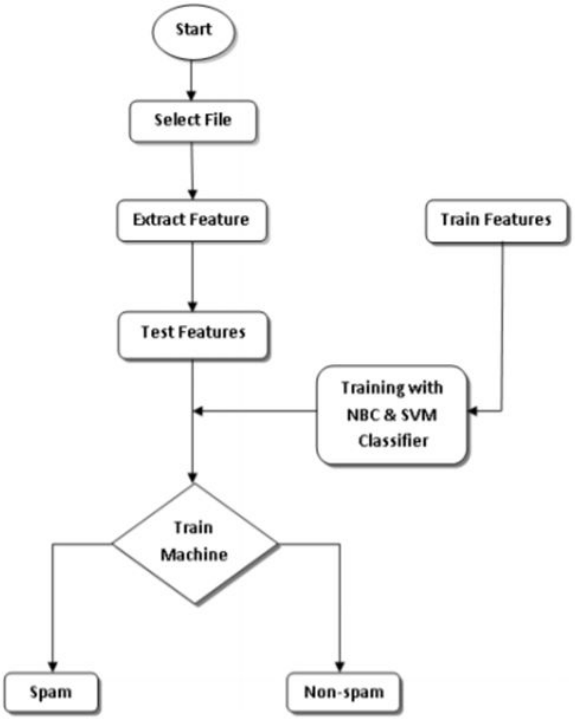
d) Finally, we have incorporated the SMS Spam Corpus v.0.1 Big. It has 1,002 SMS ham messages and 322 spam messages.

4. PROBLEM STATEMENT

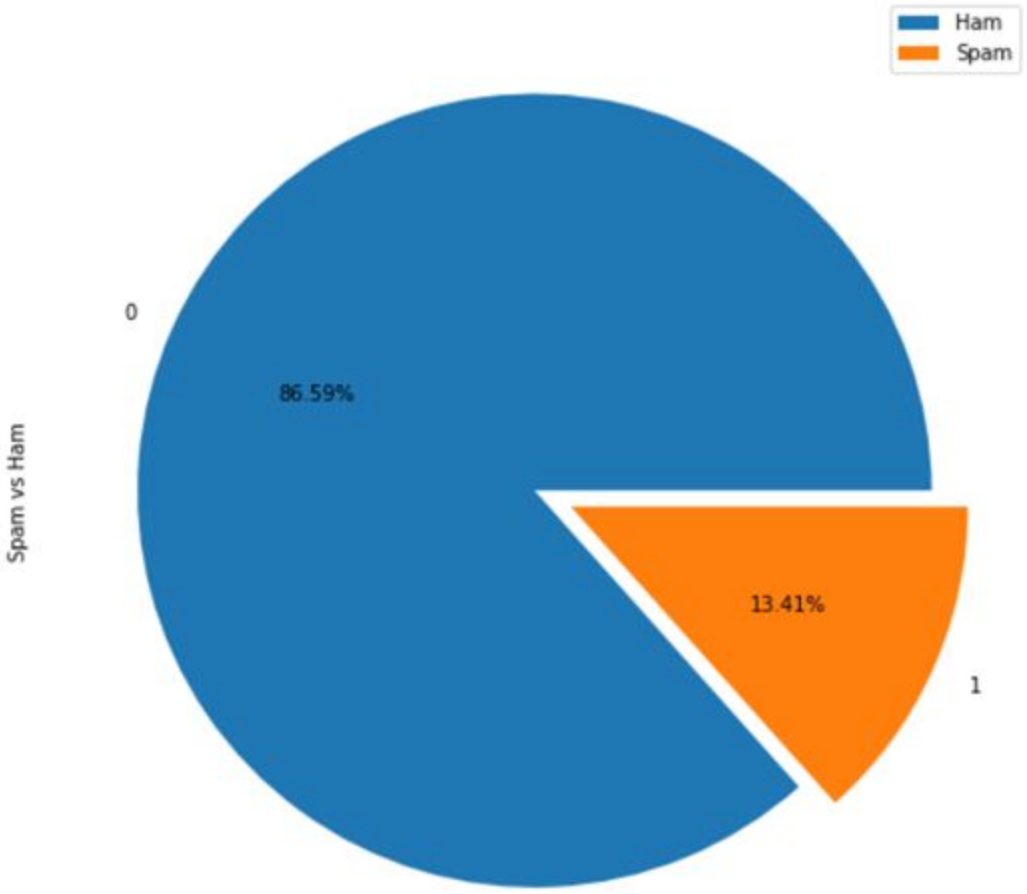
Email Spam is the most crucial matter in a social network. There are many problems created through spam. The spam is nothing; this is unwanted messages or mail which the end user does not want in our mailbox. Because of these spams the performance of the system can be degraded and affect the accuracy of the system. To send the unsolicited or unwanted messages which are also called spam is used in Electronic spamming. In this project, explain about email spam, where spam can spoil the performance of the mailing system and build a web app to detect the spam and non-spam mails.

5. SYSTEM DESIGN

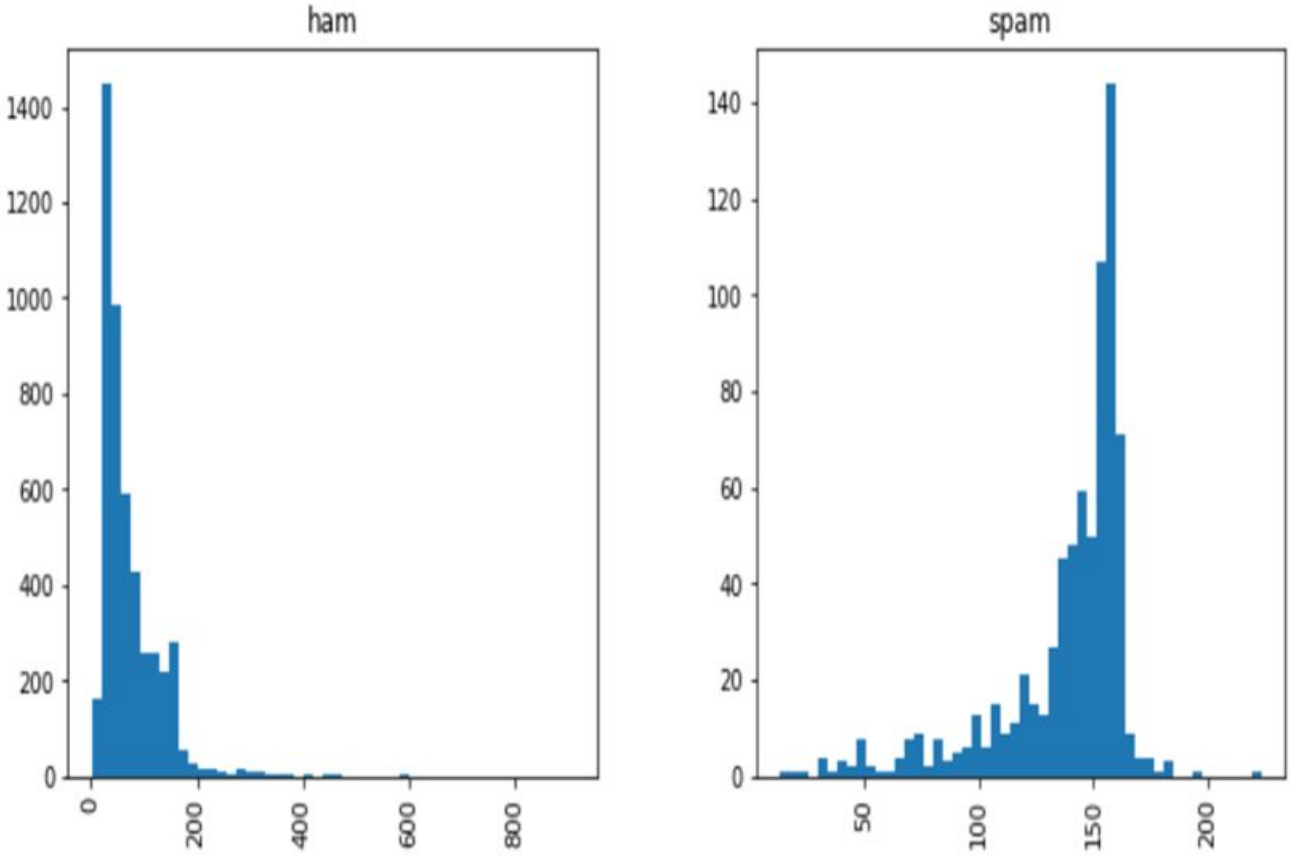
The system is built upon a micro framework named Flask using Python as a programming platform. In this work we are describing the method which is used to perform e-mail spam classification. The first step is to select the file from the dataset and apply the feature extraction technique for extracted features. For which we are using the word-count algorithm. The next step is training the dataset which is extracted by the feature extraction technique. For training the data we can calculate the probability of spam and non-spam words in the document. The next step is to test the data with the help of Binomial Naïve Bayesian Classifier for which calculation the probability of spam and non-spam mails and make a prediction which value is higher. If spam words are greater than non-spam words in a mail, then the mail is spam mails otherwise non-spam mails. In the next step we are calculating the words which are wrongly classified by the classifier and calculate accuracy of the classifier and calculate the error rate of the classifier by calculating the fraction of word which is wrongly classified and total number of words in document.



6. Data Insights



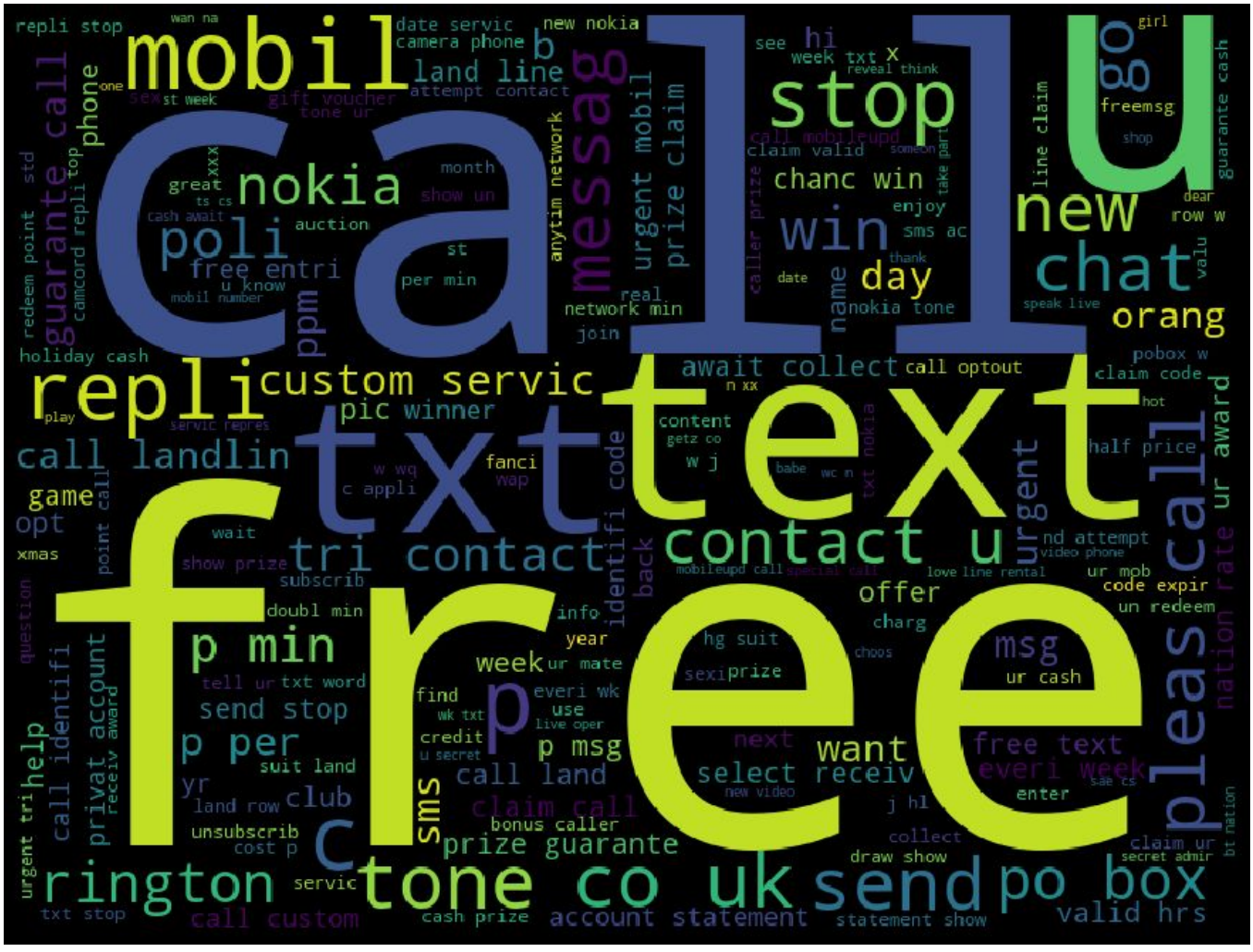
Here in this visualization, I have used the pie chart to get the percentage of spam and ham words that are there in the data set. This Pie chart shows that there are only 13.41% spam messages and 86.59% ham messages in the data set.



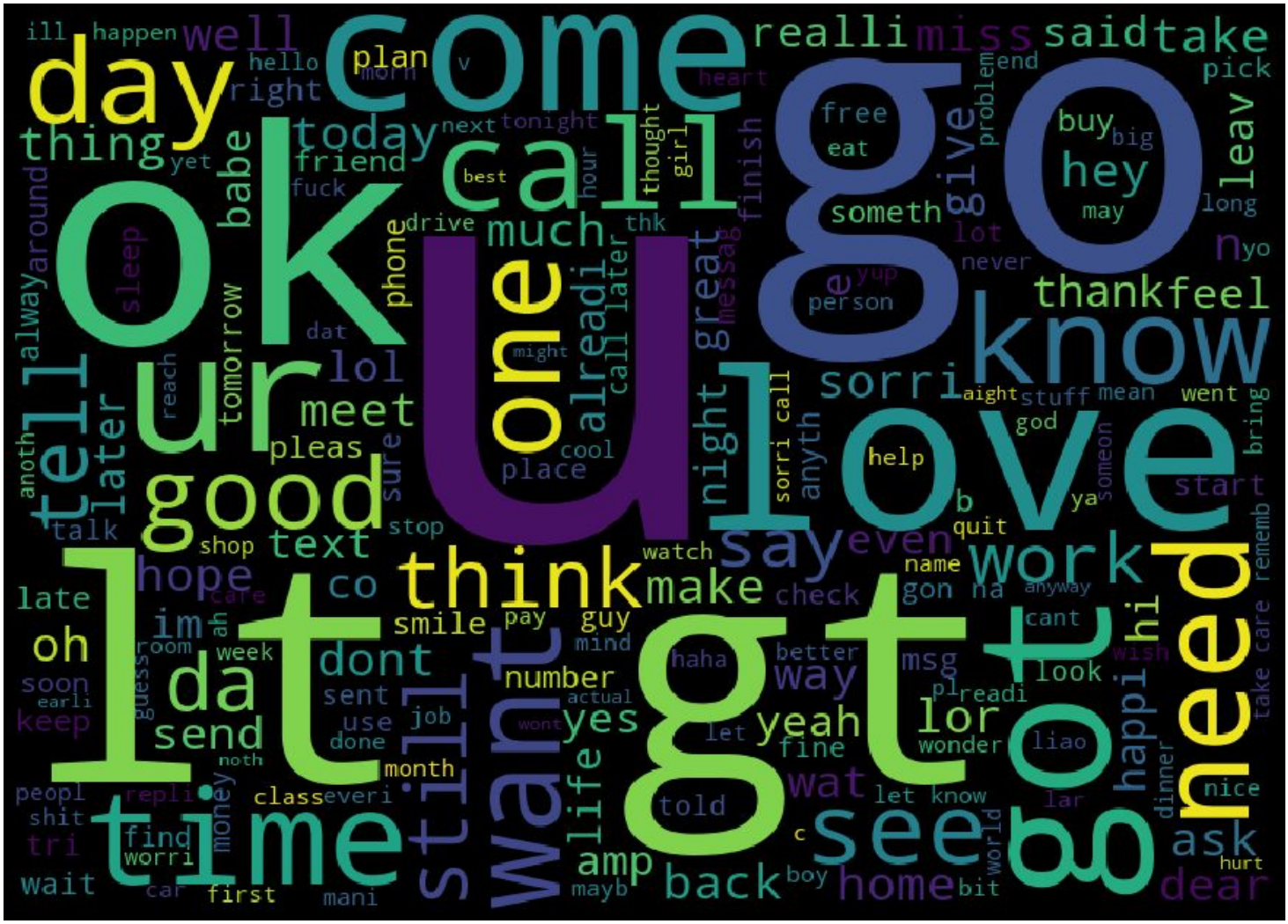
Here this visualization shows whether there is any relation between label and length. So, I have visualized the number of messages and the length of the messages.

In the Ham graph you can see that there are less number of messages based on the length of the message while on the right hand side in the Spam graph the number of messages are more based on the length of the message as compared to the left graph.

I have made word clouds for both ham and spam to do simple text analysis. Word Clouds display the most prominent or frequent words in a body of text (such as a State of the Union Address). Typically, a Word Cloud will ignore the most common words in the language (“a”, “an”, “the” etc). Below is the representation of Spam words. It is indicating all the spam words are present in the data set. Word clouds or tag clouds are graphical representations of word frequency that give greater prominence to words that appear more frequently in a source text. The larger the word in the visual the more common the word was in the document. Free, Call, Mobil, Contact, Stop these words are the most common words used in spam emails.

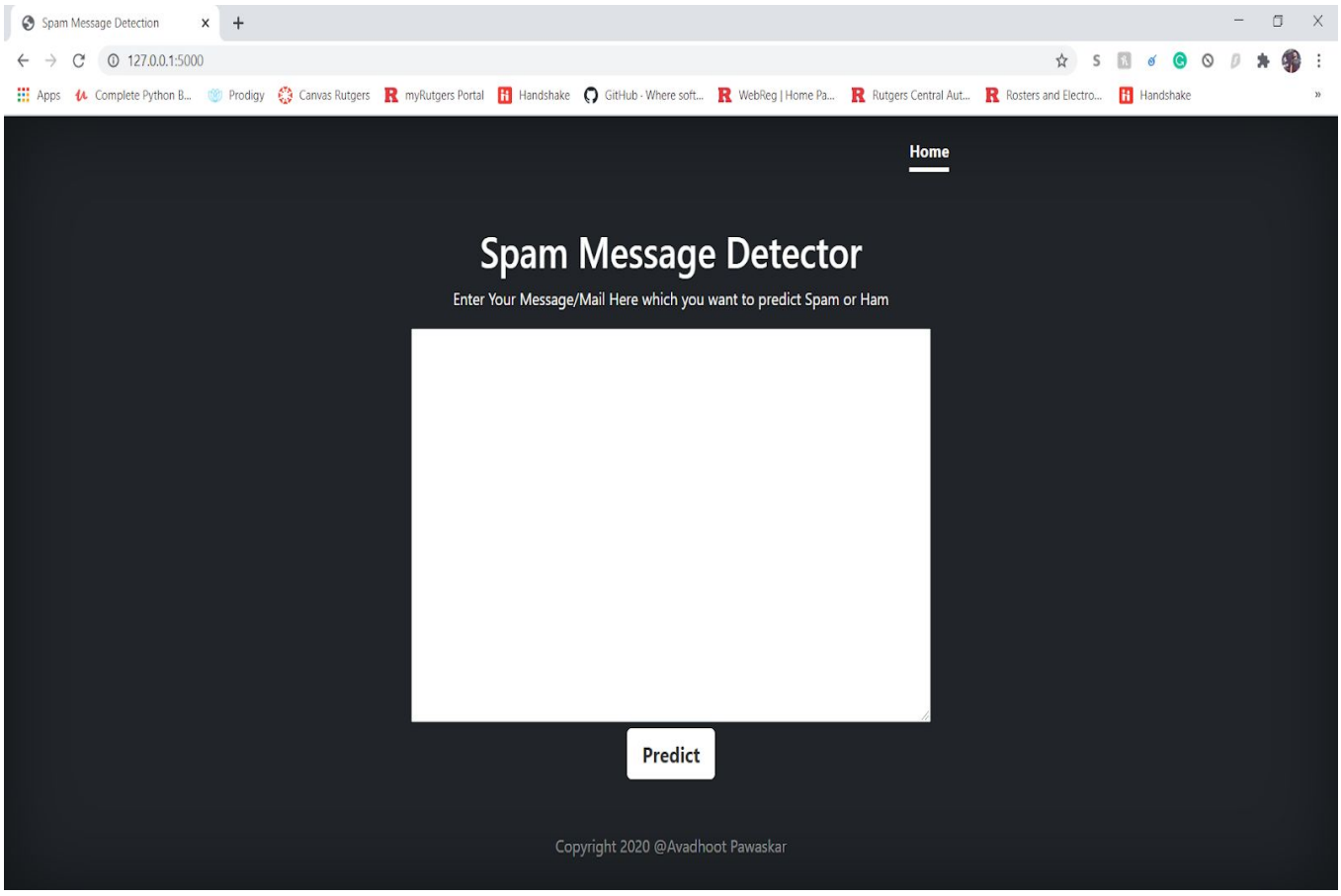


Below is the representation of Ham words. It is indicating all the spam words are present in the data set. Word clouds or tag clouds are graphical representations of word frequency that give greater prominence to words that appear more frequently in a source text. The larger the word in the visual the more common the word was in the document. Free, Come, Go, Love, Ok these words are the most common word was used in spam emails.

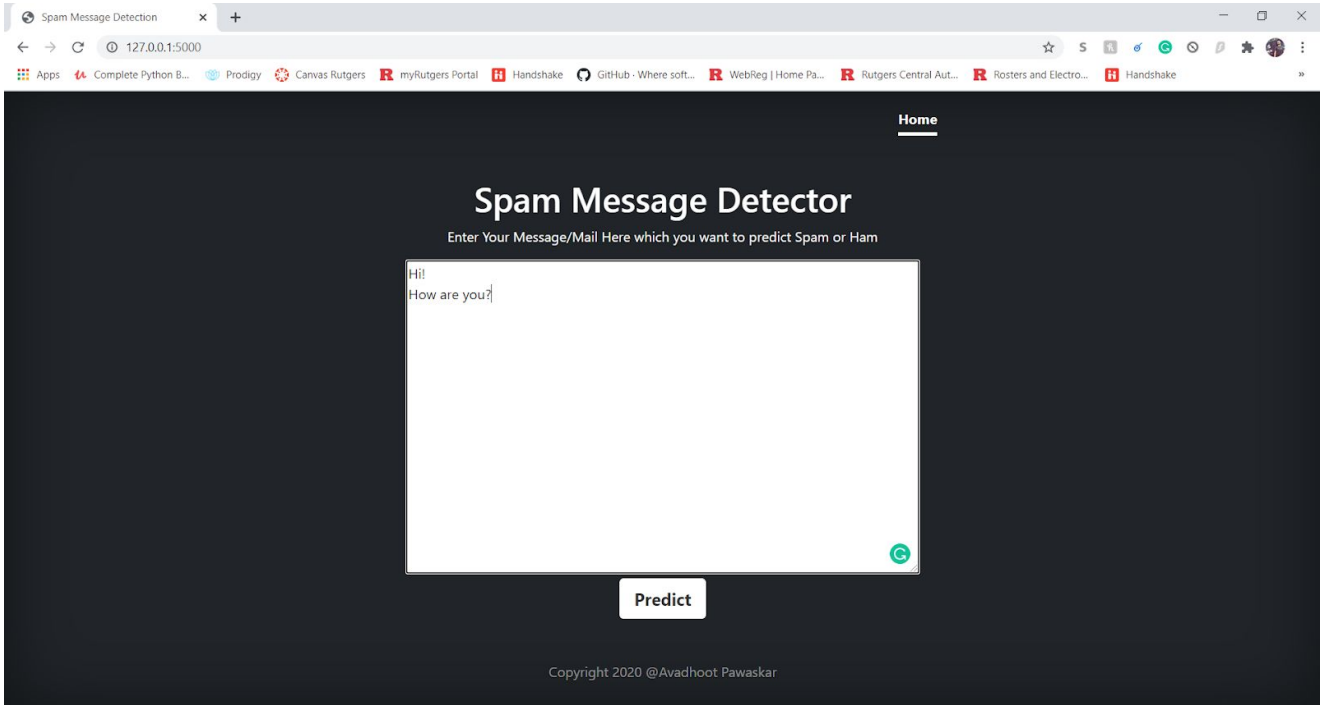


7. APP FRONTEND

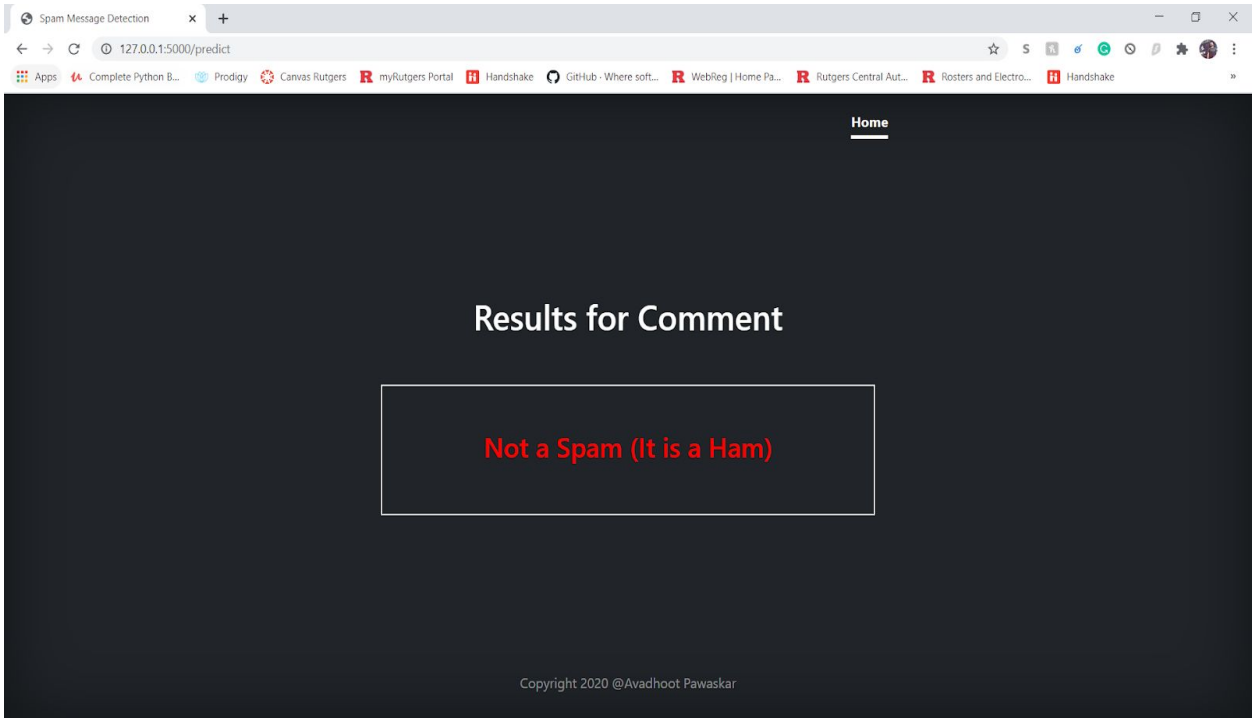
For the frontend of the web application, I have built this home page which includes Home option in the navigation bar and also there is Textbox for user to put the email message in the Text Box check whether the entered message is spam or ham (not a spam) using Predict button.



Here users have put the input in the Text Box to check whether it is a spam or not.



Based on the input given by the user, the detector has predicted the following output.



For the accuracy of the model, I have used fbeta_score as the database was imbalanced. The accuracy score I got is 93.09%.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(df, y, test_size = 0.20, random_state = 0)

# Training model using Naive bayes classifier

from sklearn.naive_bayes import MultinomialNB
spam_detect_model = MultinomialNB().fit(X_train, y_train)

y_pred=spam_detect_model.predict(X_test)
```

```
print(accuracy_score(y_test,y_pred))
print(fbeta_score(y_test,y_pred,beta =0.5))
```

```
0.9811659192825112
0.9390862944162438
```

8. CONCLUSION

In this project work we are explaining about the e-mail spam classification to identify the spam and non-spam mails. For this purpose we are using Multinomial Naïve Bayesian Classifiers. In this project we are creating an email spam classification system to classify the spam and non-spam mails. For this I have taken the collection of different datasets in one csv file called Spam.csv to run this experiment. In a dataset I have taken a total 5572 mails in which 4458 train dataset and 1114 test dataset. Out of 5572 dataset the 747 are spam mails and 4425 are non-spam mails.

The results which are provided by the classifier, we can say that the Naive Bayesian Classifier classifies mostly words in an accurate way. When the number of dataset is increased the Multinomial Naive Bayesian Classifier produces a better result as compared to other classifiers.

Spam is a big problem of today's world; to solve this problem the spam classification system is created to identify the spam and non-spam mails. The spam messages are the unwanted messages which the end user clients are receiving in our daily life. Spam mails are nothing, it is the advertisement of any company, any kind of virus etc.

To solve this problem create an email spam classification system and identify the spam and non-spam mails. Here we are using the Naïve Bayesian Classifier and extracting the word using word-count algorithm. After calculation we find that the Binomial Naïve Bayesian classifier has more accuracy than the other classifiers. The error rate is very low when we are using the Multinomial Naïve Bayesian Classifier. So we can say that Multinomial Naïve Bayesian Classifier produce better results than other classifiers.

9. References

1. https://medium.com/@engrravijain/spam-classifier-flask-app-b3ba30964c2
2. https://towardsdatascience.com/develop-a-nlp-model-in-python-deploy-it-with-flask-step-by-step-744f3bdd7776
3. http://cs229.stanford.edu/proj2013/ShiraniMehr-SMSSpamDetectionUsingMachineLearningApproach.pdf
4. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1205&context=amcis2014