

Spam Detection Project

Submitted by:

VIVEK KUMAR SINGH

**ACKNOWLEDGMENT**

I take great pleasure to thank and acknowledge the help provided by **Flip Robo Technologies**. I extend whole hearted thanks to Mr. Shwetank Mishra who become my Mentor and with whom I worked and learned a lot and for enlightening me with her knowledge and experience to grow with the corporate working. Her guidance at every stage of the Project enabled me to successfully complete this Project which otherwise would not have been possible without her consent encouragement and motivation. Without the support it was not possible for me to complete the report with fullest endeavour.

**INTRODUCTION**

* Business Problem Framing

Email is the worldwide use of communication application. It is because of the ease of use and faster than other communication application. However, its inability to detect whether the mail content is either spam or ham degrade its performance. Nowadays, lot of cases have been reported regarding stealing of personal information or phishing activities via email from the user. This project will discuss how machine learning help in spam detection. Machine learning is an artificial intelligence application that provides the ability to automatically learn and improve data without being explicitly programmed. Binary classifier will be used to classify the text into two different categories; spam and ham. The algorithm will predict the score more accurately. The objective of developing this model is to detect and score word faster and accurately.

* Conceptual Background of the Domain Problem

Today, spam has become a big internet issues. Recent 2017, the statistic shown spam accounted for 55% of all e-mail messages, same as during the previous year. Spam which is also known as unsolicited bulk email has led to the increasing use of email as email provides the perfect ways to send the unwanted advertisement or junk newsgroup posting at no cost for the sender. This chances has been extensively exploited by irresponsible organizations and resulting to clutter the mail boxes of millions of people all around the world. Evolving from a minor to major concern, given the high offensive content of messages, spam is a waste of time. It also consumed a lot of storage space and communication bandwidth. End user is at risk of deleting legitimate mail by mistake. Moreover, spam also impacted the economical which led some countries to adopt legislation. 2 Text classification is used to determine the path of incoming mail/message either into inbox or straight to spam folder. It is the process of assigning categories to text according to its content. It is used to organized, structures and categorize text. It can be done either manually or automatically. Machine learning automatically classifies the text in a much faster way than manual technique. Machine learning uses pre-labelled text to learn the different associations between pieces of text and it output. It used feature extraction to transform each text to numerical representation in form of vector which represents the frequency of word in predefined dictionary. Text classification is important in the context of structuring the unstructured and messy nature of text such as documents and spam messages in a cost-effective way. A Machine learning platform has capabilities to improve the accuracy of predictions. With regard to Big Data, a Machine Learning platform has abilities to speed up analysing of gigantic data. It is important especially to a company to analyse text data, help inform business decisions and even automate business processes. For example, text classification is used in classifying short texts such as tweets or headlines. It can be used in larger documents such as media articles. It also can be applied to social media monitoring, brand monitoring and etc. In this project, a machine learning technique is used to detect the spam message of a mail. Machine learning is where computers can learn to do something without the need to explicitly program them for the task. It uses data and produce a program to perform a task such as classification. Compared to knowledge engineering, machine learning techniques require messages that have been successfully pre-classified. The pre-classified messages make the training dataset which will be used to fit the learning algorithm to the model in machine learning studio. 3 A specific algorithm is used to learn the classification rules from these messages. Those algorithms are used for classification of objects of different classes. The algorithms are provided with input and output data and have a self-learning program to solve the given task. Searching for the best algorithm and model can be time consuming. The two-class classifier is best used to classify the type of message either spam or ham. This algorithm is used to predict the probability and classification of data outcome.

* Review of Literature

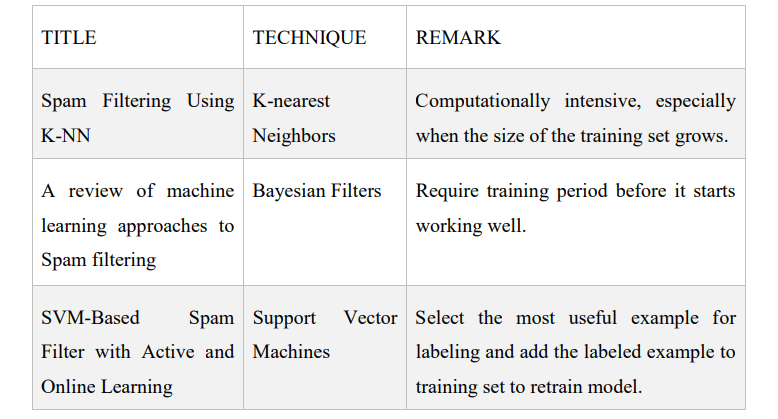
This chapter discusses about the literature review for machine learning classifier that being used in previous researches and projects. It is not about information gathering but it summarizes the prior research that related to this project. It involves the process of searching, reading, analysing, summarising and evaluating the reading materials based on the project. Literature reviews on machine learning topic have shown that most spam filtering and detection techniques need to be trained and updated from time to time. Rules also need to be set for spam filtering to start working. So eventually it become burdensome to the user.

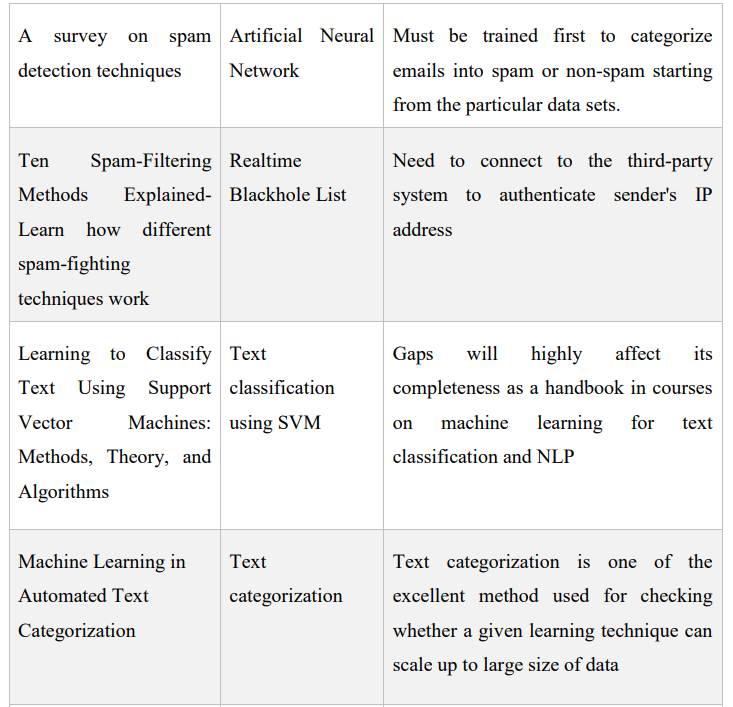
MACHINE LEARNING In this project, existing machine learning algorithm is used and modified to fit the need of project. The reasons are because machine learning algorithm is adept at reviewing larges volume of data. It is typically improves over time because of the ever-increasing data that are processed. It gives the algorithm more experience and be used to make better predictions. Machine learning allows for instantaneous adaption without human intervention. It identifies new threats and trends and implements the appropriate measures. It is also save time as it is it automated nature.

SPAM DETECTION In theory, spam detection can be implemented at any location and multiple stages of process can occur at the same time.

Most research has been conducted into detecting and filtering spam email using a variety of techniques. Thiago S. Guzella et. Al (2009) has conducted “A Review of Machine Learning Approaches to Spam Filtering”. In their paper, they found that Bayesian Filters that are used to filter spam required a long training period for it to learn before it can completely well function. S. Ananthi (2009) has conducted a research on “Spam Filtering using K-NN”. In this paper, she used KNN as it is one of the simplest algorithm. An object is classified by a majority vote of its neighbours where the class is typically small. Anjali Sharma et. Al (2014) has conducted “A Survey on Spam Detection Techniques”. In this paper, they found that Artificial Neural Network (ANN) must be trained first to categorize emails into spam or non-spam starting from the particular data sets. Simon Tong and Daphne Koller (2001) has conducted a research on “Support Vector Machine Active Learning with Applications to Text Classification”. In this paper, they presented new algorithm for active learning with SVM induction and transduction. It is used to reduce version space as much as it can at every query. They found out that the existing dataset only differ by one instance from the original labelled data set. 10 Minoru Sasaki and Hiroyuki Shinnou (2005) has conducted a research on “Spam Detection Using Text Clustering”. They used text clustering based on vector space model to construct a new spam detection technique. This new spam detection model can find spam more efficiently even with various kinds of mail. Aigars Mahinovs and Ashutosh Tiwari (2007) has conducted a research on “Text Classification Method Review”. They test the process of text classification using different classifier which is natural language processing, statistical classification, functional classification and neural classification. They found that all the classifier works well but need more improvement especially to the feature preparation and classification engine itself in order to optimise the classification performance.

Table 1 below shows the summary of techniques used by other researchers of spam detection and filtering based on different books and journals.

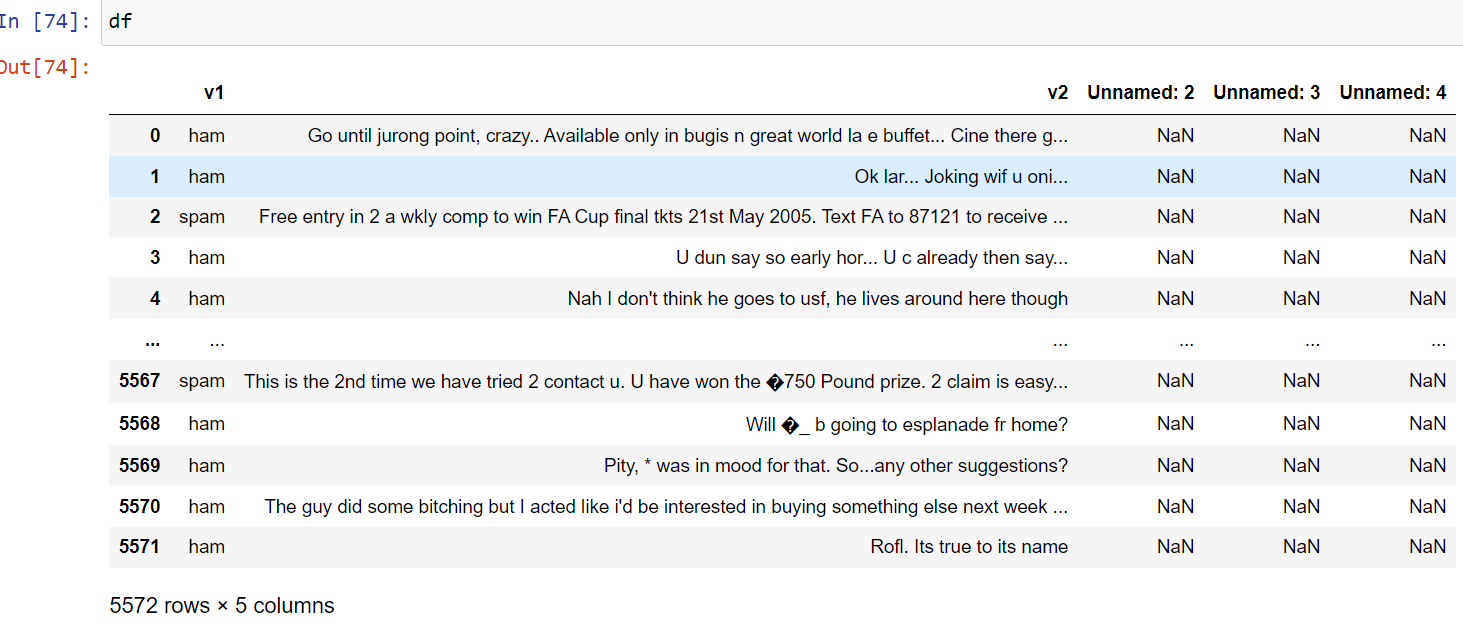




**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

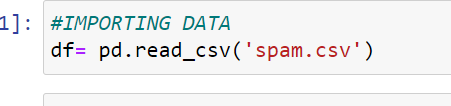
The SMS Spam Collection is a set of SMS tagged messages that have been collected for SMS Spam research. It contains one set of SMS messages in English of 5,574 messages, tagged according being ham (legitimate) or spam.



Shape of the data is 5572 rows and 5 columns .

* Data Sources and their formats

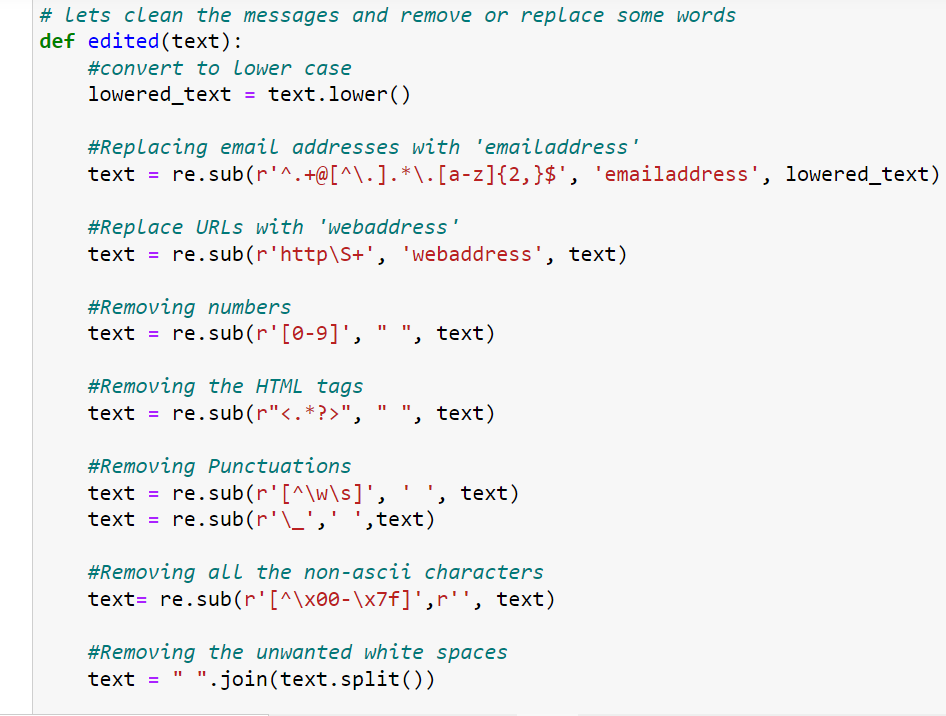
Data is provided by the company Fliprobo and format of the data is in csv format.



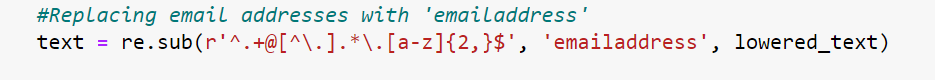
.

* Data Pre-processing Done

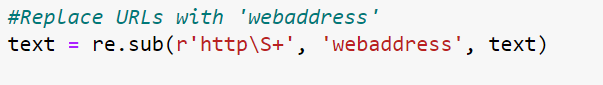
Data is cleaning and pre-processing:-



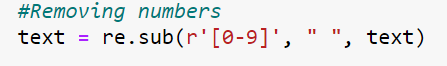
Here , I took some steps to clean and pre- processed the dataset. First it converted into lower text, so that all the dataset words should be in lower text. I have changed the email address with “emailadress”.



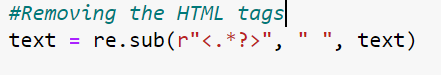
Replacing URLs with “webaddress”.



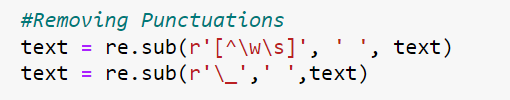
Removing numbers.



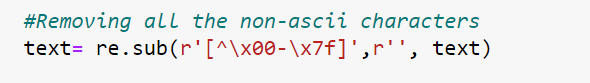
Removing html tags.



Removing punctuations



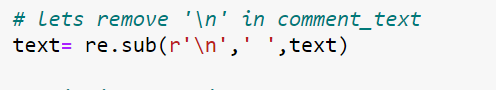
Remove non-ascii characters



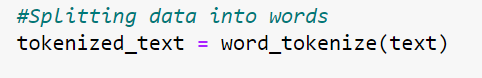
Removing white space

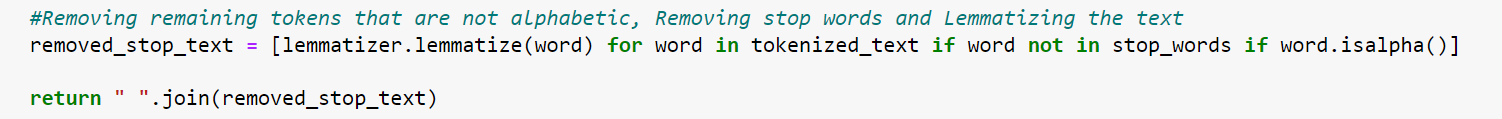


Removing “\n”



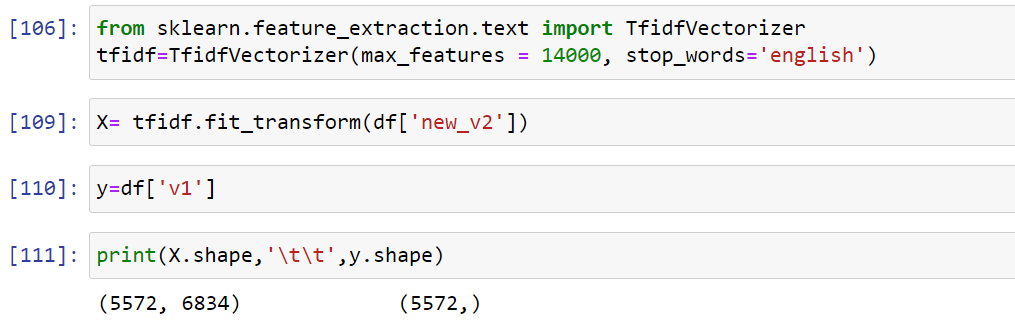
Splitting into words





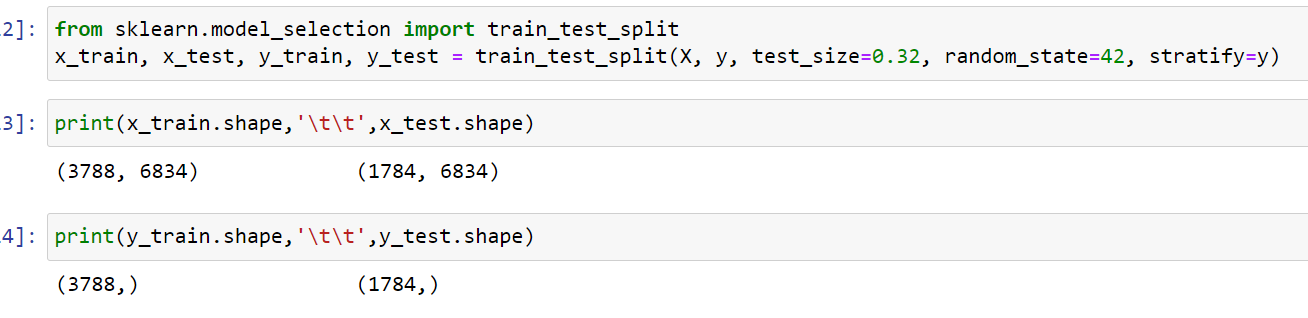
so, these are the steps taken for cleaning the data.

PRE-PROCESSING:-



Here, I used the TFidf vectorizer to convert it into algorithm for model training and prediction.

* Data Inputs- Logic- Output Relationships



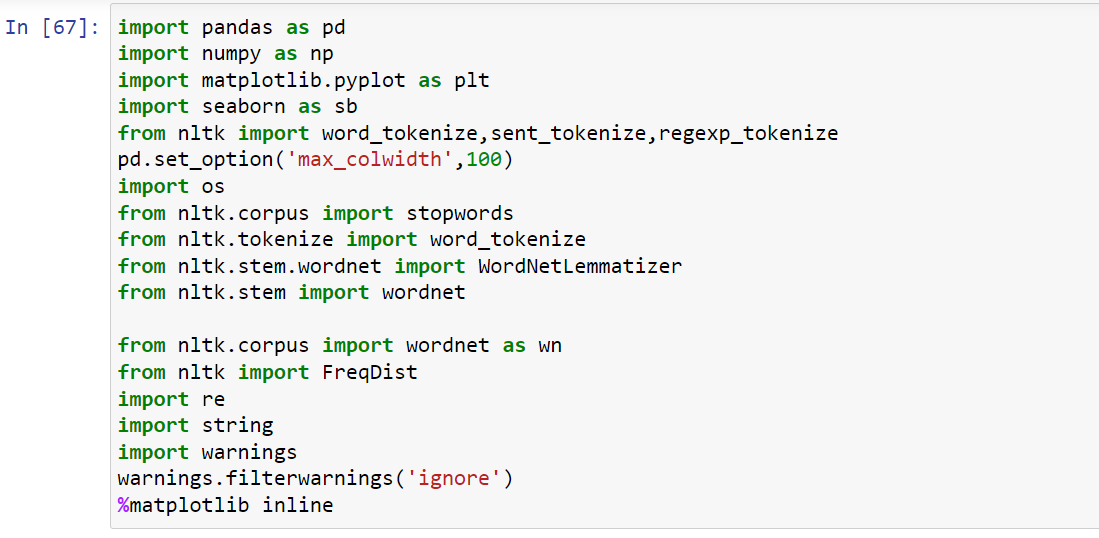
We split x-train,x-test and y-train ,y-test for trainig and testing the model.

First it train and we feed x\_test as input data and predict and compare the y\_test which is actual output.

* Hardware and Software Requirements and Tools Used

The project is done into laptop with i5 processor with quad core with 8gb of RAM with GTX 1650 GPU with Anaconda and juypter notebook.

Some of important packages is imported in python for this projects which are



.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

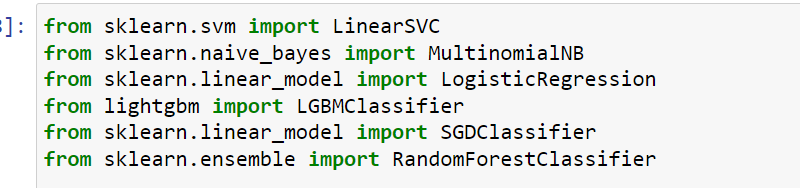
As per the problem it is the supervised learning problem so, I used here supervised learning apporches.

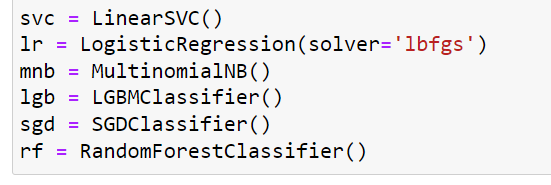
Problem solving approaches are :-

* Removing special character
* Removing emojis and emails
* Removing punctuation
* Removing stop words
* Removing numbers.
* Using lemmatize and stemming
* Using vectorizer (tfidf vectorizer) for numerical values
* Splitting train and test data
* And at the end create models
* Check the accuracy
* Use cross validation for overfitting and underfitting
* Saving the best performance model

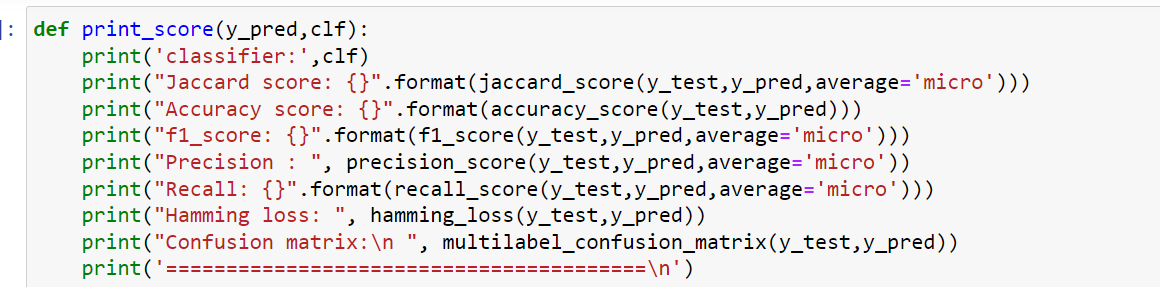
* Testing of Identified Approaches (Algorithms)

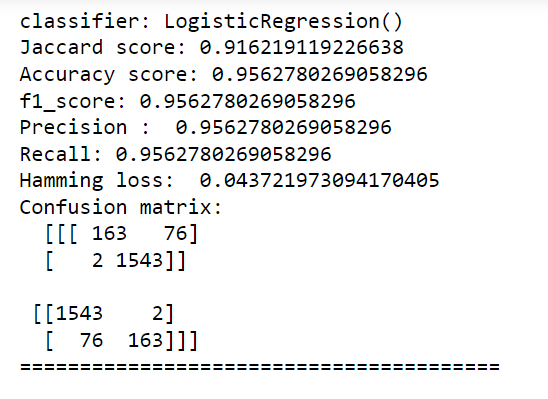
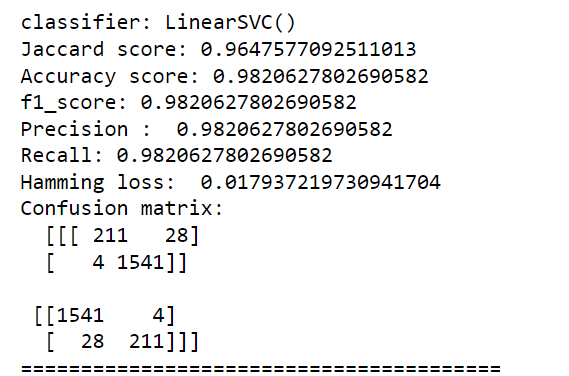
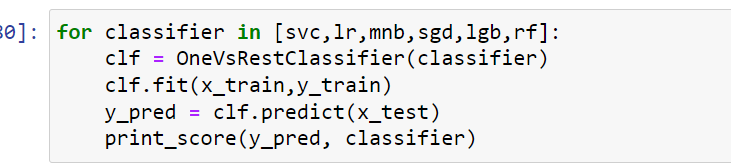
The following algorithm is used for testing .

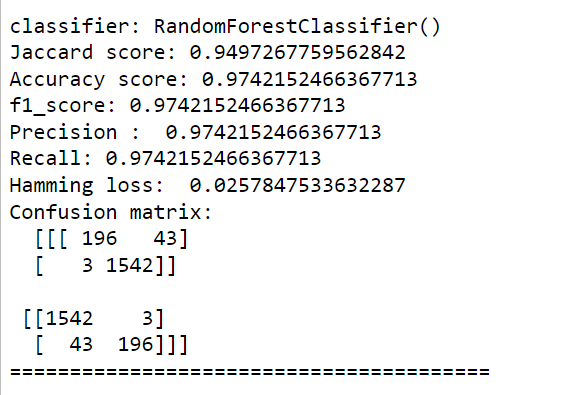
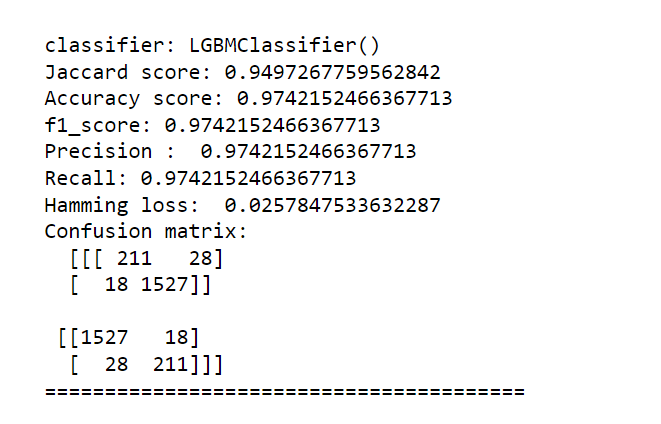
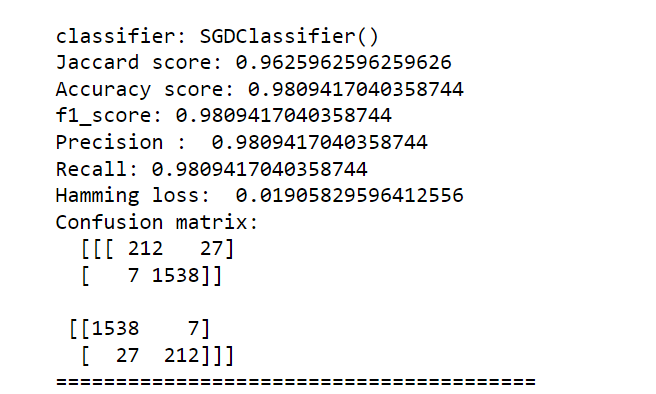
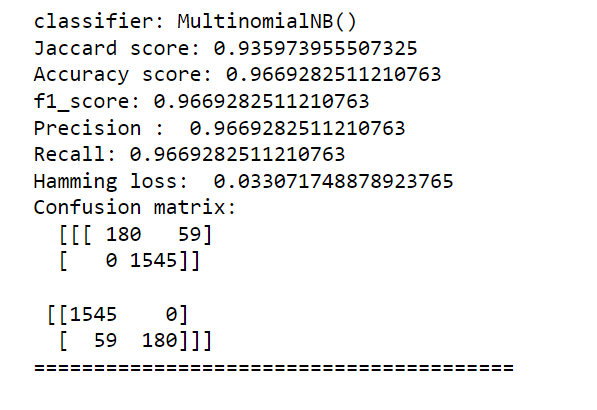




* Run and Evaluate selected models



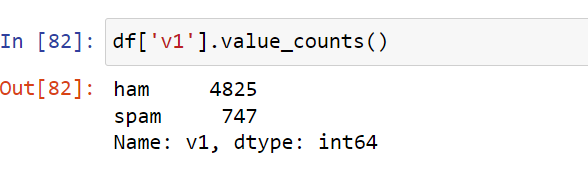


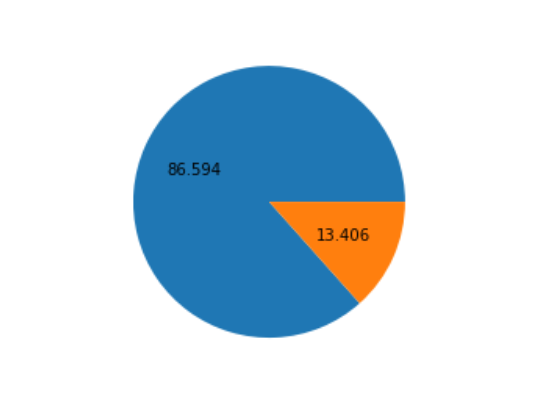
.

* Key Metrics for success in solving problem under consideration
* In basis of accuracy score, matrix and cross validation. Logistic regression selected as final models.

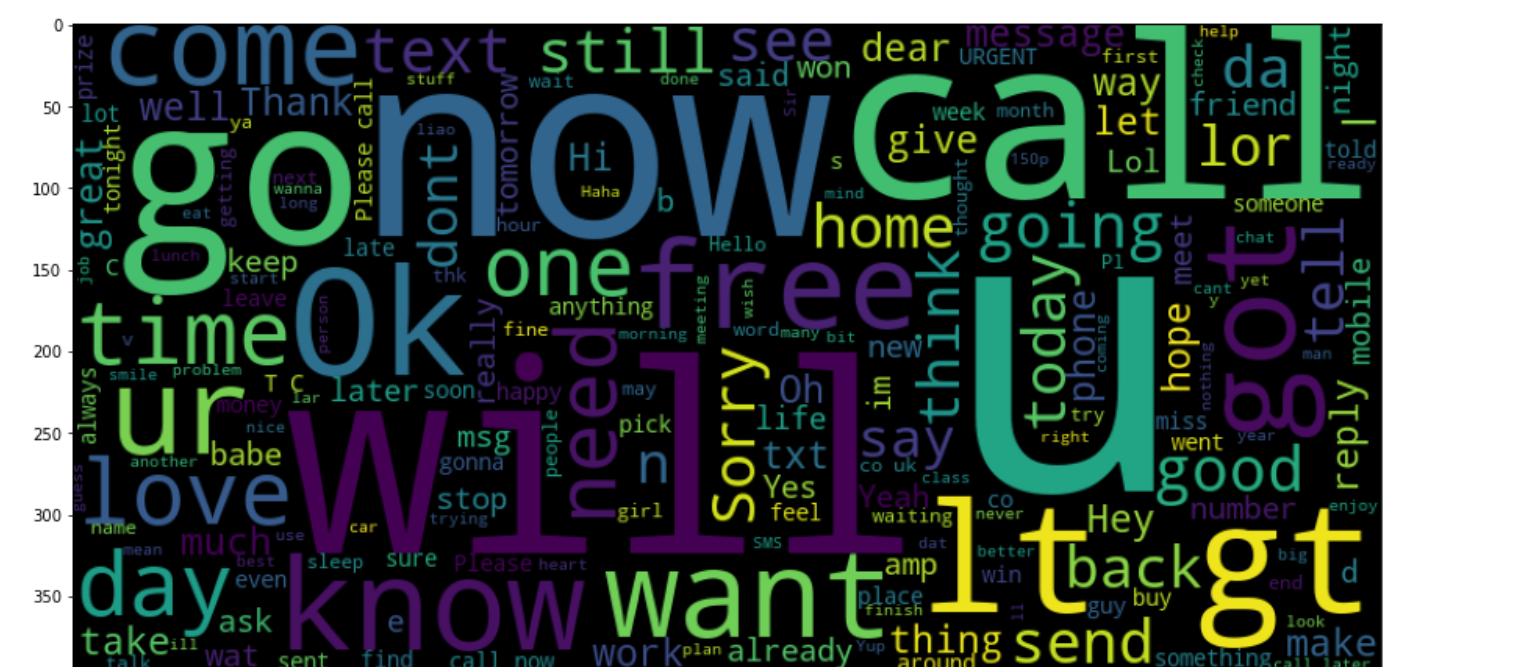
.

* Visualizations

this is the target variable ham and spam.



This is how the ratio look like 86% ham and 13% spam. Although it is very clear that the data is unbalanced here.



This is the figure before cleaning the data.



In this figure this all the words we see is the spam data , this is all spam data words.

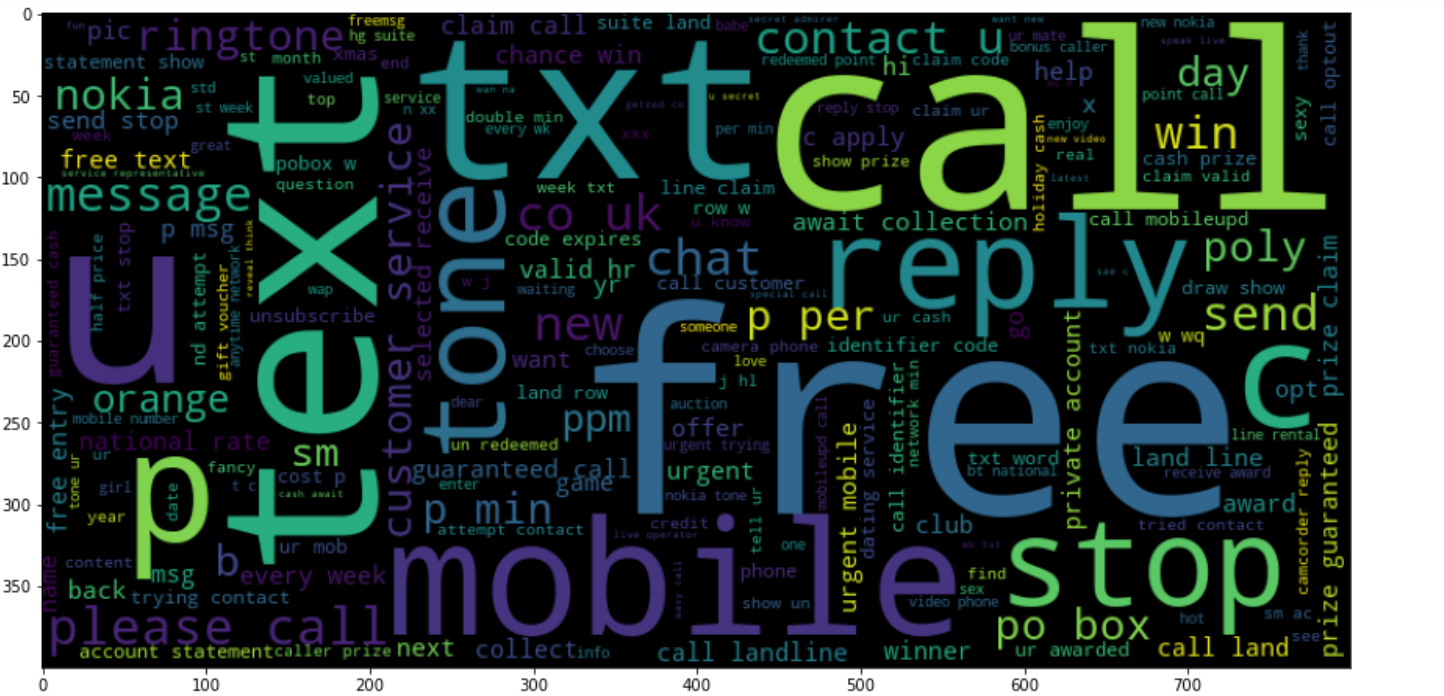


In this figure we see all ham data words.

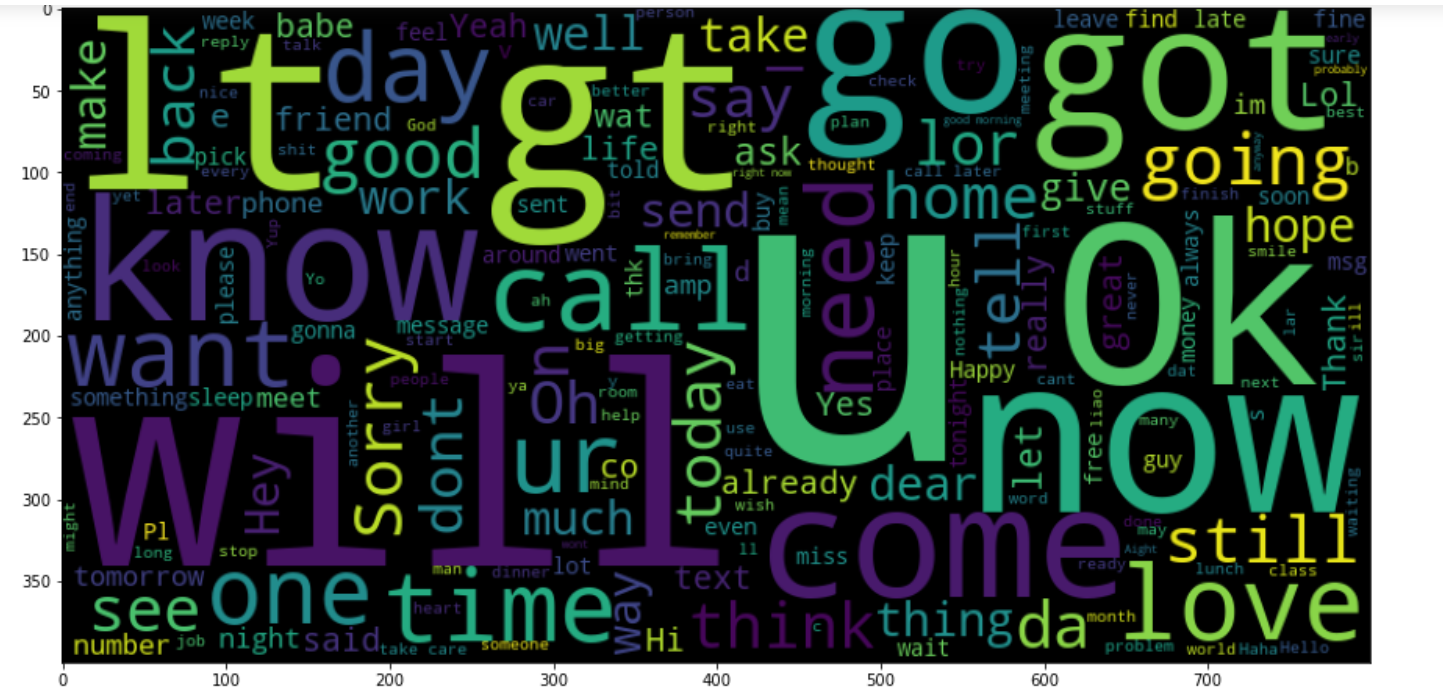
Here these figure are before cleaning the data. Lets see how data look like after the cleaning it.



This is whole data here spam and ham both are here.



This figure has the spam data after cleaning.



This is ham data after the cleaning the data.

**CONCLUSION**

The performance of a classification technique is affected by the quality of data source. Irrelevant and redundant features of data not only increase the elapse time, but also may reduce the accuracy of detection. Each algorithm has its own advantages and disadvantages as stated in . As state before, supervised ML is able to separate messages and classified the correct categories efficiently. It also able to score the model and weight them successfully. For instances, Gmail’s interface is using the algorithm based on machine learning program to keep their users’ inbox free of spam messages. During the implementation, only text (messages) can be classified and score instead of domain name and email address. This project only focus on filtering, analysing and classifying message and do not blocking them. Hence, the proposed methodology may be adopted to overcome the flaws of the existing spam detection.