SPAM DETECTION

Submitted by:

ANJANA.P

ACKNOWLEDGMENT

I would like to express my appreciation to team Fliprobo for giving such a data for analysis, with a full-length description of the project. My mentor Mr.Sajid Choudhary has helped me in many stages of this project where I was stuck with problems. I use this opportunity to thank him for helping me at the right time without any delay.

Also, this project made me search for a lot of data's in several webpages and sites, that helped me to rectify my doubts and, I was able to study more about NLP.

INTRODUCTION

We were asked to assume that we are hired in a start-up company and are assigned with a new project. Project is that we need to build a system to identify ham and spam messages from email. We were also given a set of data which contains email .

The dataset has three columns named subject, message and also the label column.

It is a basic Natural Language Processing project based on sentiment analysis. There are two types of mails. They are ham and spam mails. Now we have to classify them as ham and spam using machine learning and NLP techniques.

DATA PRE-PROCESSING

First the important libraries for preprocessing and also csv file for analysis is imported. Shape and datatypes of columns are checked. There are 2893 rows and 3 columns in our dataset.. Unnecessary columns should be dropped as they provide no necessary information for our analysis. Ham and Spam value counts are checked.There are 2412 ham emails and 481 spam emails in our dataset.Then a new column is added to the dataset for the length of the message.
In preprocessing main steps include removal of punctuations and special formats like url,phone number etc and also converting all cases to lower.

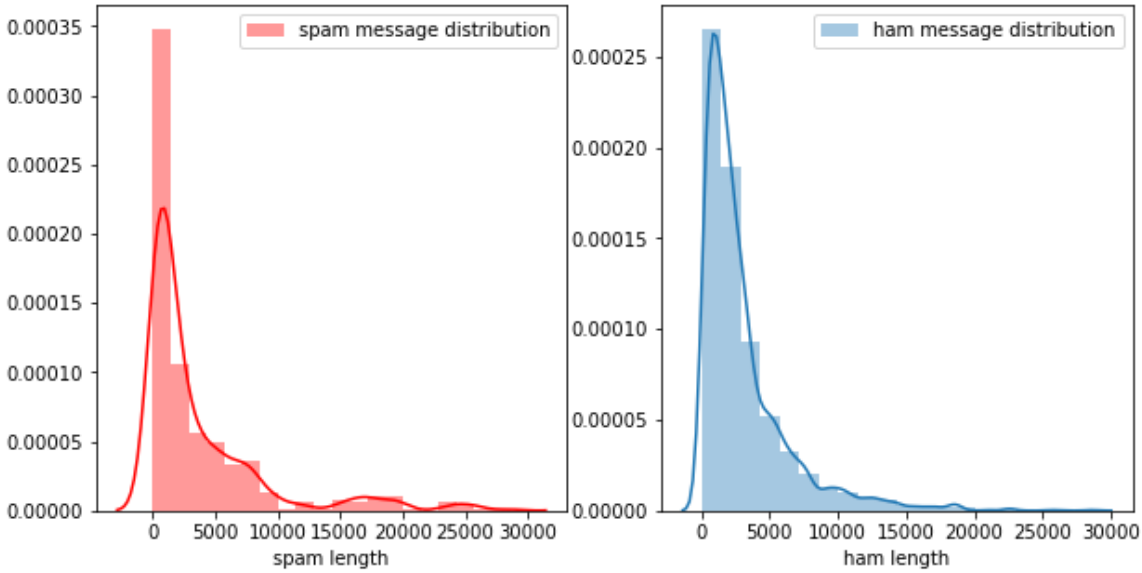
1) NLTK library-stopwords are downloaded from nltk library.Stopwords are English words

Which doesnot add much meaning to a sentence.stopwords are also removed from our dataset .List of stopwords can be found in corpus module.To remove the stopwords from a data first divide the text into words and then remove the word if it exists in the list of stopwords provided by NLTK.

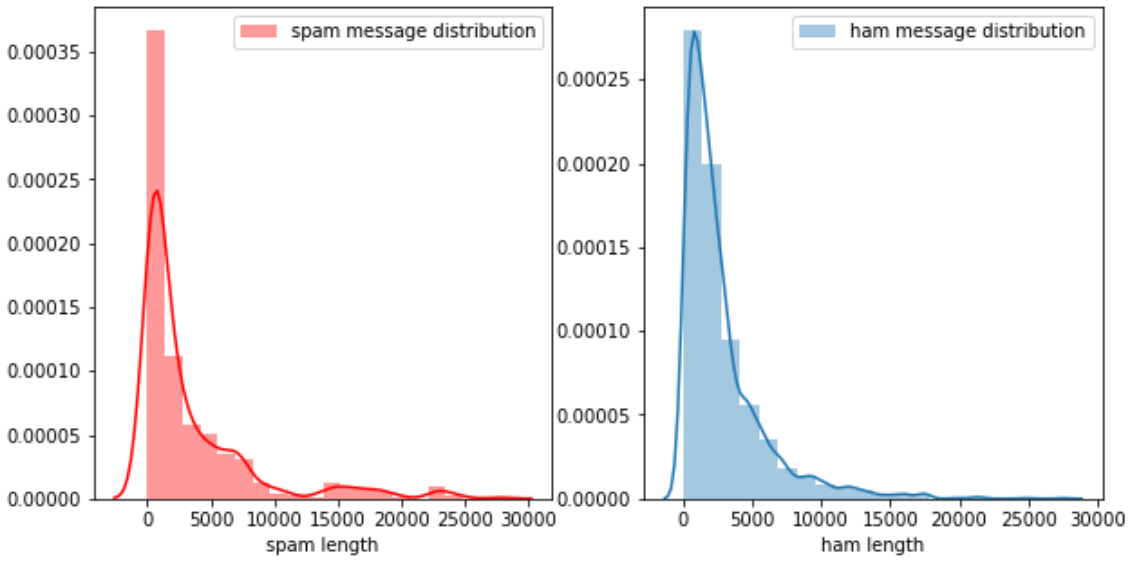
Porter stemmer- stemmers remove morphological affixes from words leaving only the word stem.It extracts the base of a modified word.so the efficiency of any content based spam filter can be significantly improved.
Stopwords
Corpus- it is a language resource consisting of large and structured set of texts.

After cleaning the whole data a new column is added to get the new clean length of each message.

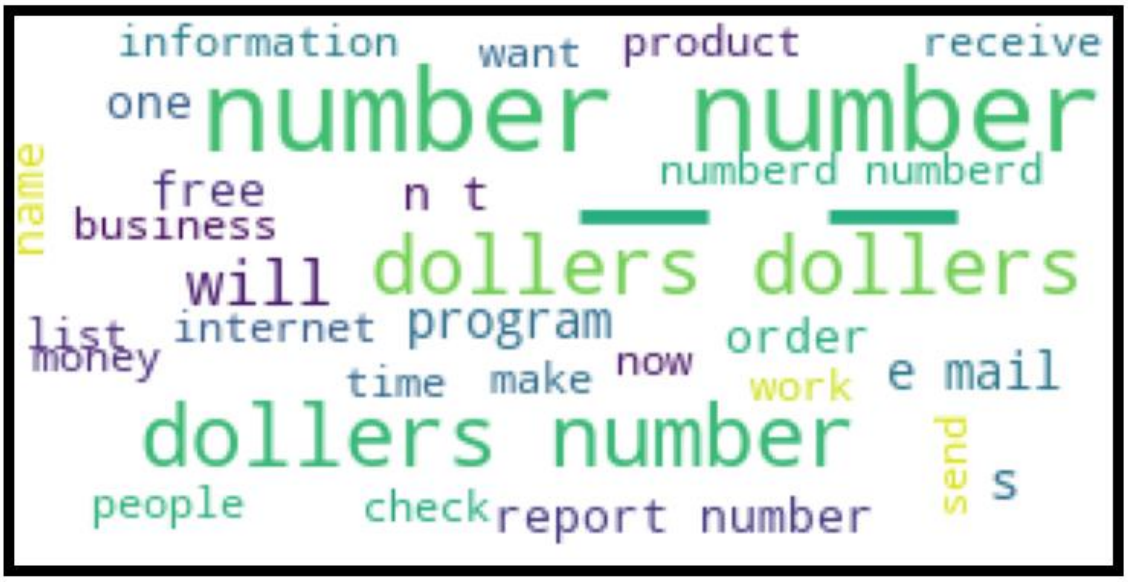
MESSAGE DISTRIBUTION BEFORE CLEANING



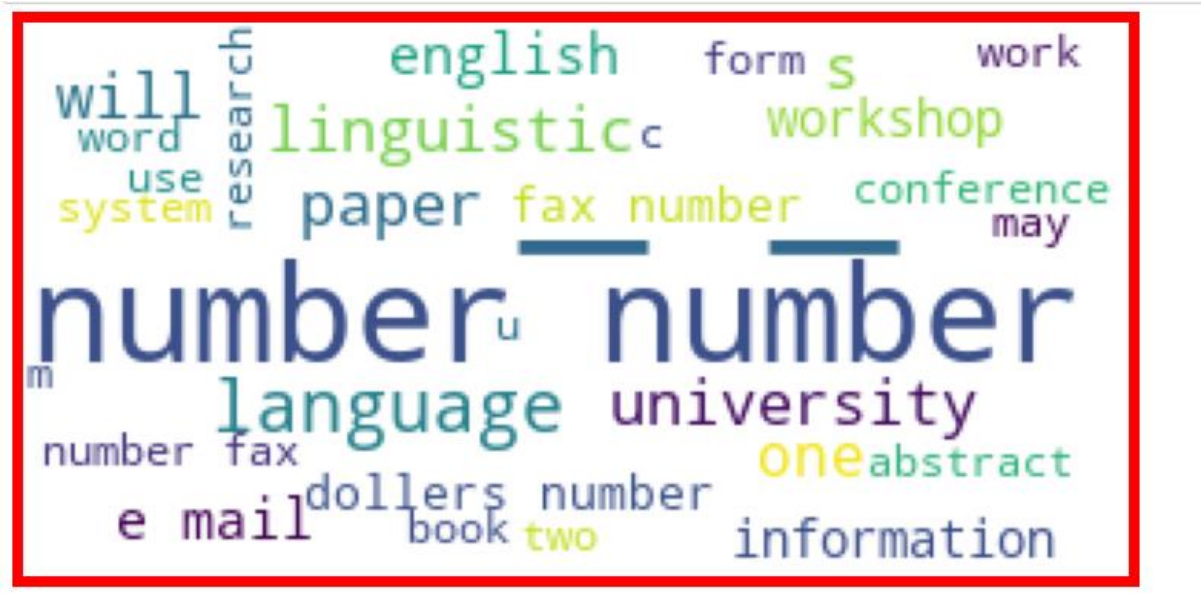
MESSAGE DISTRIBUTION AFTER CLEANING



LOUD WORDS IN HAM AND SPAM



Loud words are the most frequent words seen in a data. Above figure shows the loud words in spam. There the most frequent words seen are number, dollars , information etc.



These are the loud words in ham message. The most frequent words seen in ham messages are number , language ,university etc..

CONVERTING TEXT INTO VECTOR

TF-IDF vectorizer- It is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. A tf-idf vectorizer transform text to feature vectors that can be used as input to estimator.In each vector the number or weight represent features tf-idf score.After converting into vectors train and test datas are separated

Testing of Identified Approaches (Algorithms)

After the train test split of the data models like decision tree classifier ,auc_roc curve and decision tree curve are checked. Decision tree classifier gives the best accuracy score, so the cross-validation score of dtc is checked.

Evaluation of selected models

1)Logistic regression-It is a classification algorithm used to predict the probability of categorical dependent variable.

Logistic regression is giving an accuracy score of 0.96.

2)Multinomial Naïve Bayes-It is a variant Naïve Bayes that follows Multinomial normal distribution and supports continuous data.

Multinomial NB is giving an accuracy score of 0.83

3)RandomForest classifier-It gives an accuracy of 0.99

4)Decision tree classifier –It gives an accuracy of 0.97

5)AUC_ROC –It is an important evaluation metrics to check any models performance.auc_roc graph is drawn with different threshold valued outputs based on true positive rate and false positive rate. The more the area is under the curve, it shows that the model performs well.

We get an auc_score of 0.88

Decision tree classifier – A tree structure is constructed that breaks the dataset into smaller subsets eventually resulting in prediction.The root node partitions the data based on most influential feature partitioning.There are two measures for this. They are gini impurity and Entropy.

Random forest classifier gives a maximum accuracy of 0.99,So its cross validation score is checked.Cross_val_score of random forest classifier is 0.97

So the model is saved.