Aslı Kurt/18030411057 & 18.05.2021

Erdicem Yücesan/15030411052

**The Report Of the Spam Classification App ML**

For the begining of this report our goal is briefly consist of explanation of the Machine Learning (ML) and how can machine learning be used for the security side with an Open Source example.

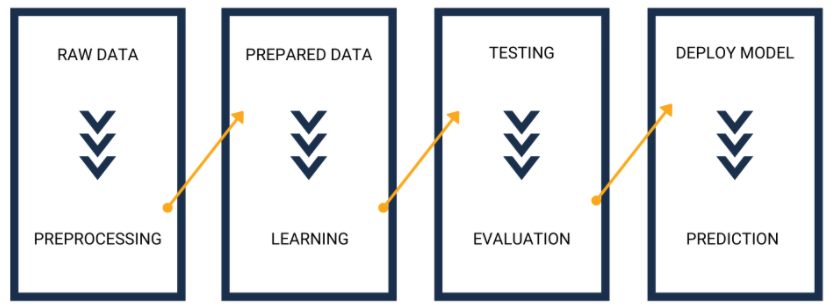
**So, What is Machine Learning?**

“Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves.” (Expert.ai Team, 6 May 2020)

Also we can say that, the real means behind ML is ‘Machine Teaching’. We know that what the machine needs to learn, so our job is to create a learning framework and provide properly formatted, relevant, clean data for the machine to learn from. (lexalytics, 2020)

**How is it used for cybersecurity?**

“Machine learning helps automate the process of finding, contextualizing, and triaging relevant data at any stage in the threat intelligence lifecycle.”(Echosec Systems, 26 May 2020)

These are the steps of the data process in the ml.

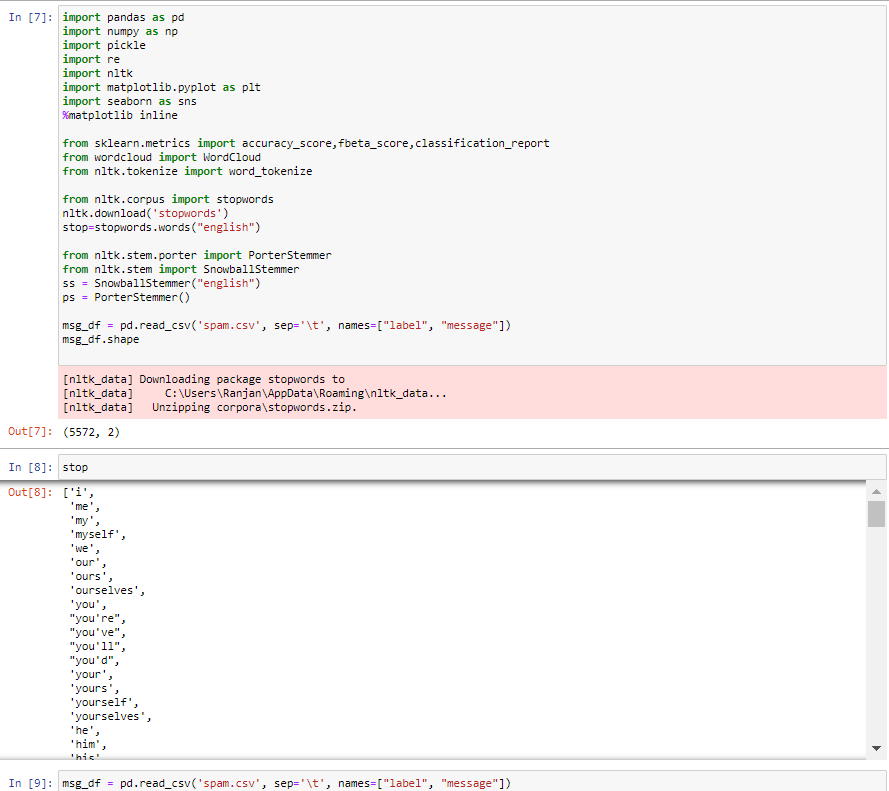
(https://www.echosec.net/blog/how-is-machine-learning-used-in-cybersecurity)

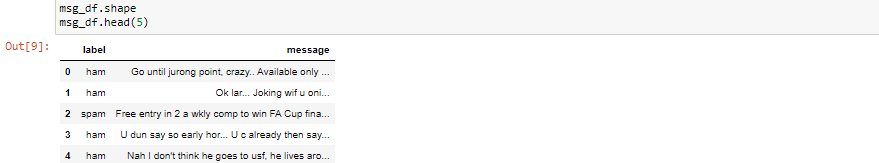
While performing the steps of the code in the open source example, the procedure given above is used.

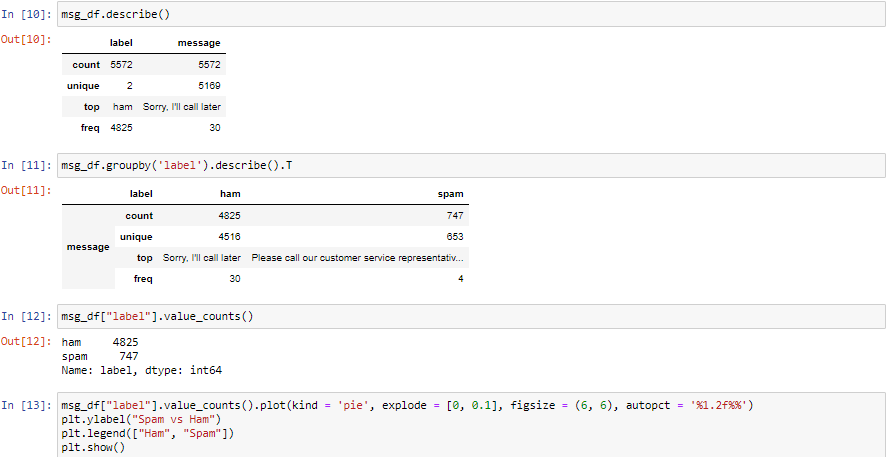
Complete Deployment of Machine Learning Model Using NLP: Spam Classifier: Modelling Part

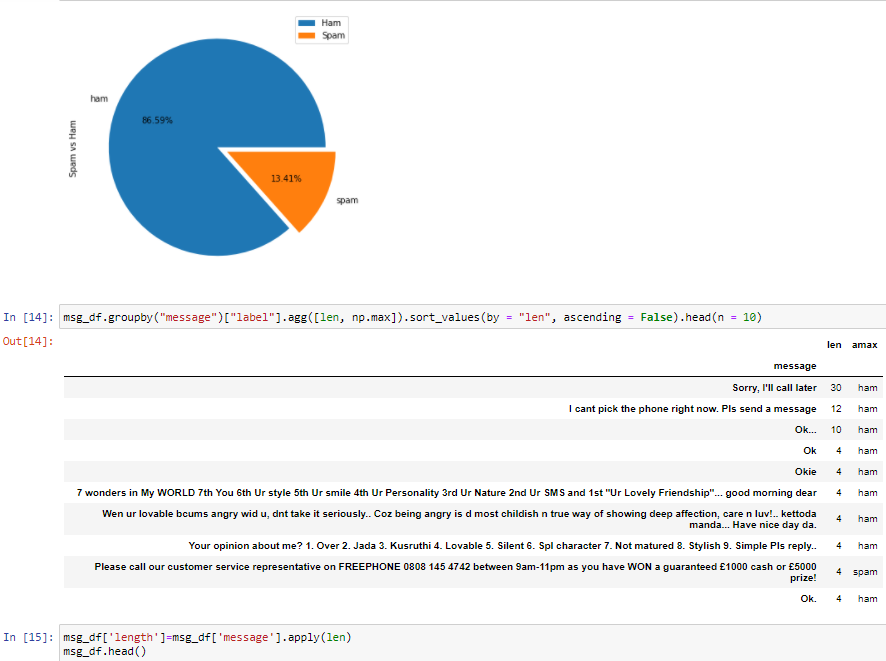
“First we create a model, once we have created a model, we will create a pickle file from the model and use that pickle file and we will deploy on the cloud herouku in order to do that we are gonna use flask framework in python But im going to show you creation of the model part”. What does this app?It checks whether received mail or message spam or not.

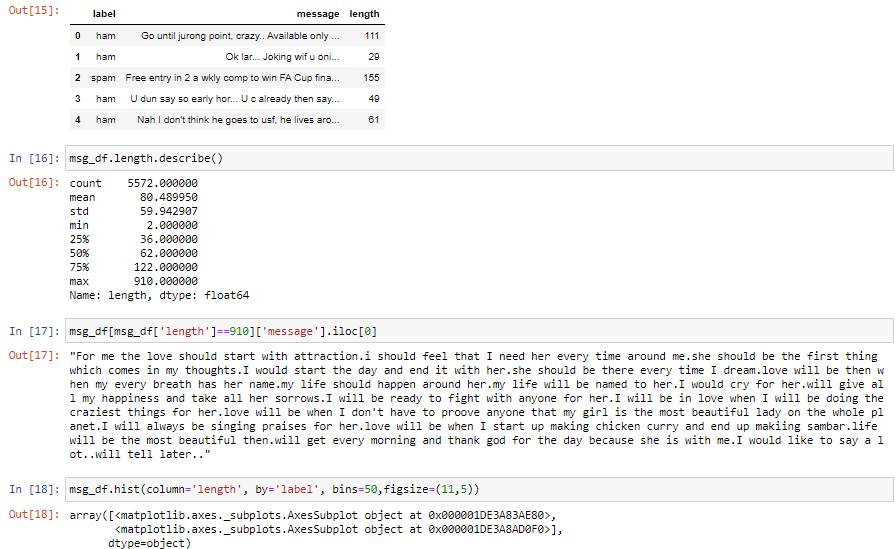
In the bellow part, there are codes that are written in the jupyter notebook and then, there is the explanation of the code.

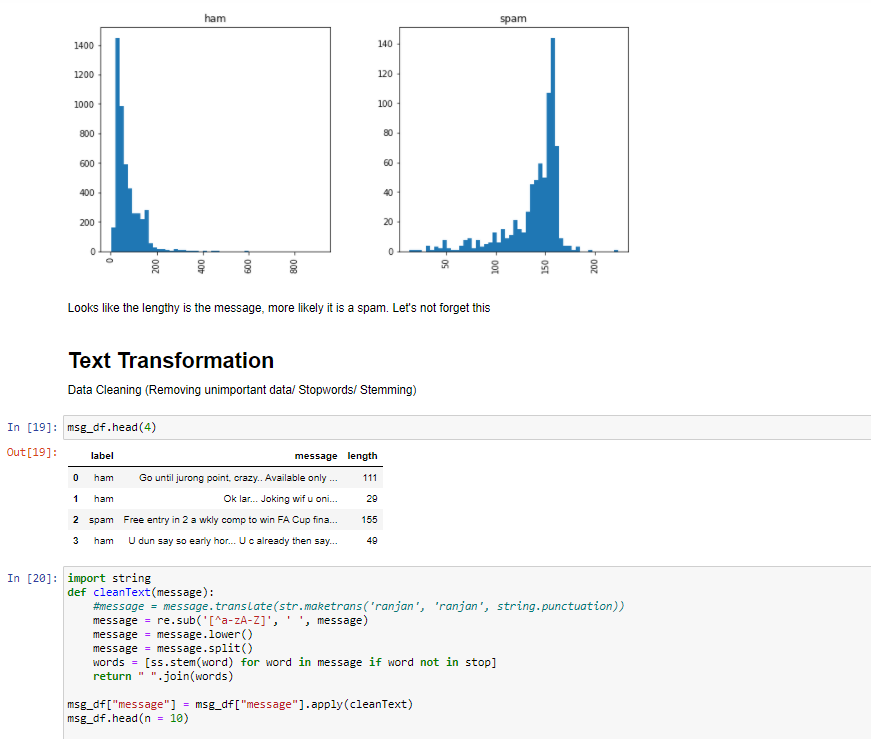


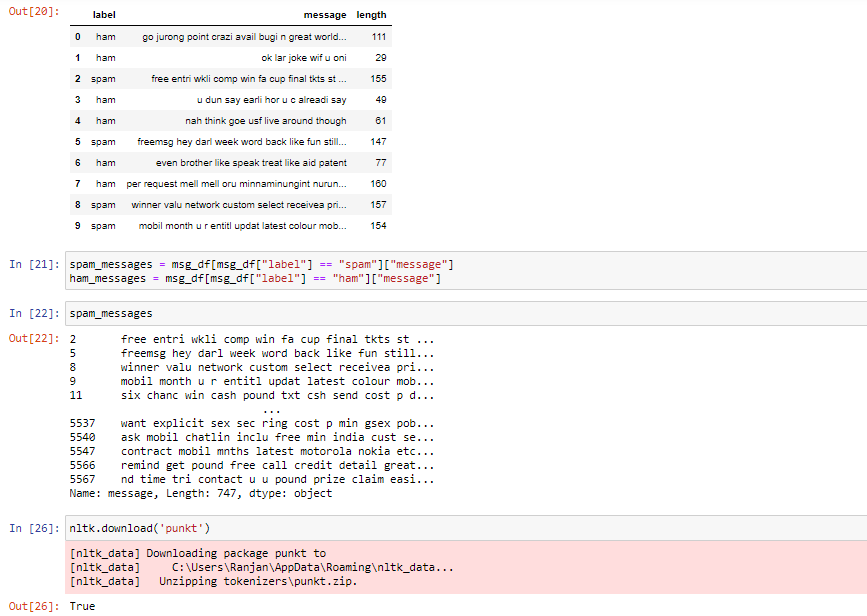


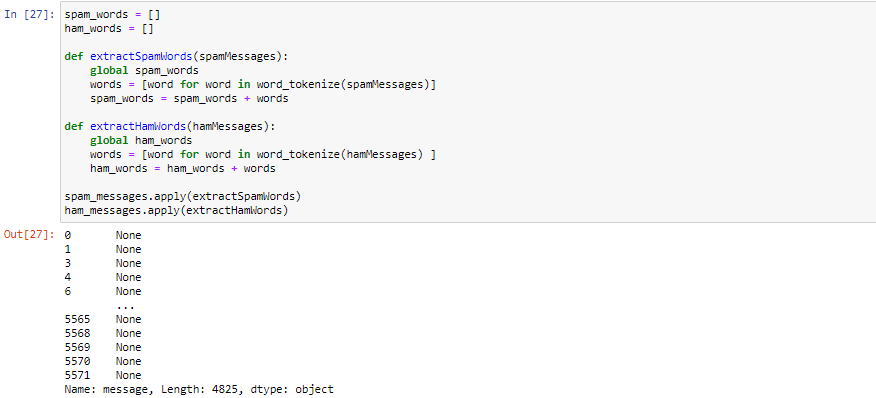




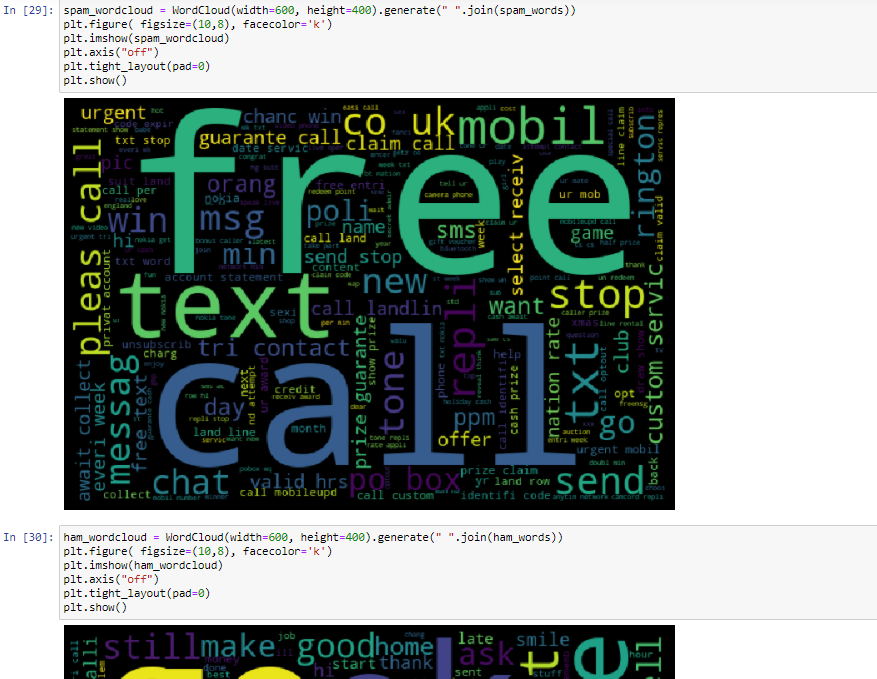


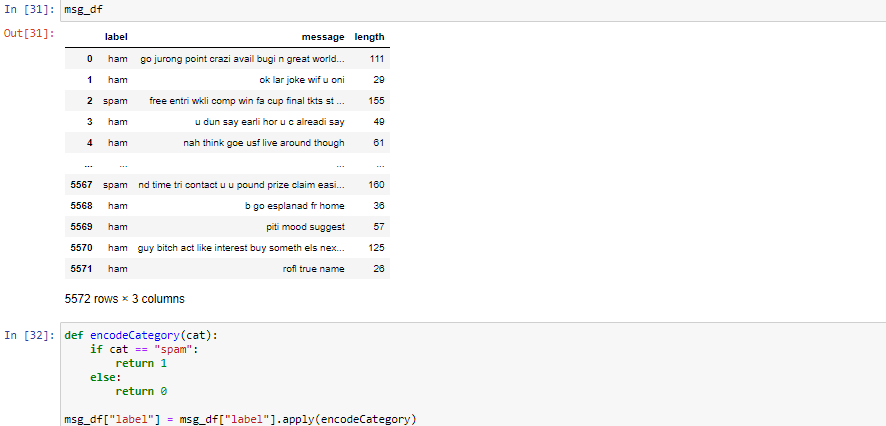


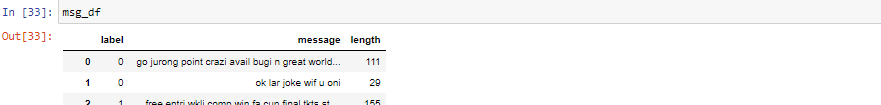






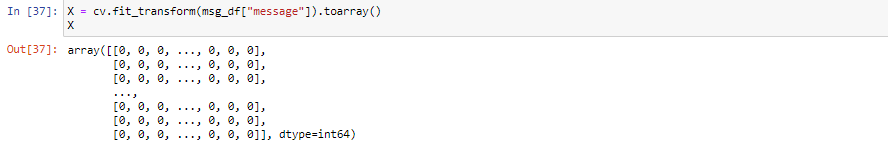


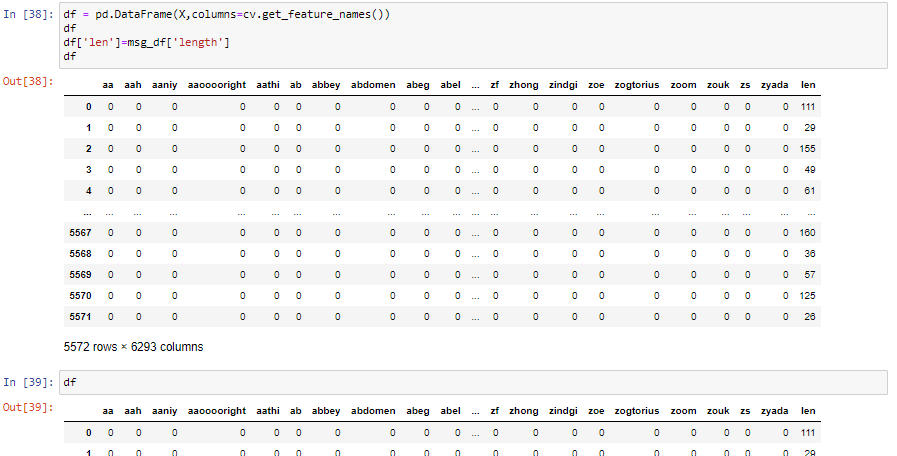


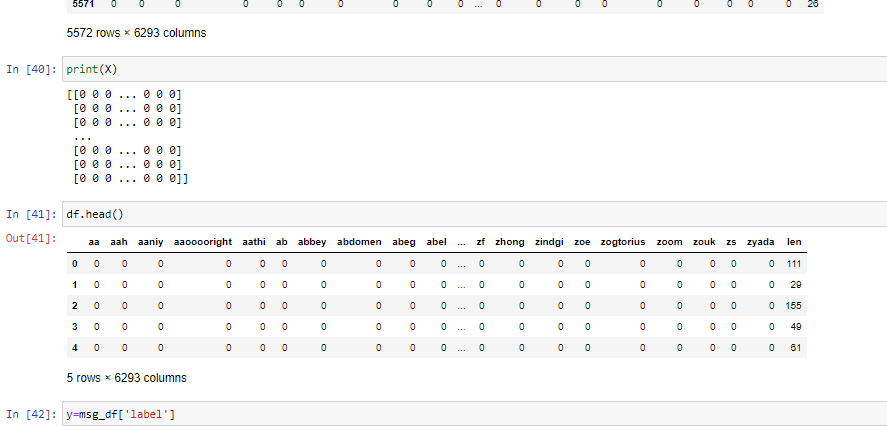


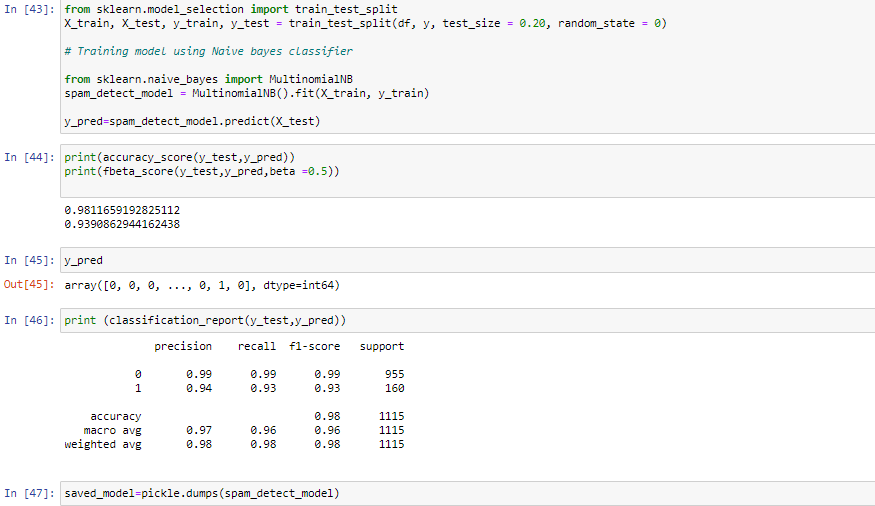


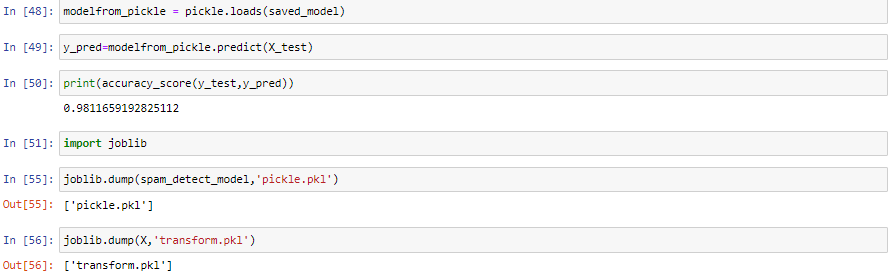












Firstly in the code, necessary libraries in python have imported such as pandas, numpy, pickle(which he will create a pickle file), re (regular expressions), nltk (natural language toolkit), matplotlib.pylot (is a collection of fuctions), seaborn (data visualization library based on matplotlib).

Then,the Accuracy score, fbeta score and classification score, they are all form the classification metrics from the sklearn.metrics.

Wordcloud imported from Word cloud which is gonna be created in the folloving sections.

word tokenizer imported from NLTK and its used for to converting dataframe for better text understanding in machine learning applications.

Stopwords basically; words that can be safely ignored withouth sacrifacing the meaning of the sentence for eg. I, We, You, They etc. and imported from nltk.corpus. If it is the first time of using this library it needs to be downloaded and then stop variable have been created and its gonna be used for storing all the stop words in english.

After that, PorterStemmer & SnowballStemmer are imported and they are basicaly extracting the words into the base/stem form(for eg.converting 3 similar words into one eg. hard, hardest and harder to hard).

new dataframe has created with the name of the msg\_df and spam.csv File has been imported(and this file consist of spam and ham messages) and then the columns have named as label (for spam/ham) and message.

The file size is five thousan five hundred seventy two with 2 columns.

These are the stop words.

Then the first five files are shown in the graph format. After that under the label and message columns, count, unique, top and freq have described.

Caunt is the total, unique has two labels which are spam & ham and unique message number is 5169, The majority consists of the ham category and the message is “Sorry. I’ll call you later.”. Since the gap between ham & sapm so big such as the number of ham is 4825 and spam is 747 then the pie chart with pyplot has been made in order to see in visualized form.

Since here it has 2 labels and now, the third label which is length has been added. The length label basicly counts the words in the message. After that, the first five messages have been shown with length. Then, length instances have been shown, such as total message, mean, standard deviation, min, max, etc and we can clearly see that, the max message length is 910. Since the message length is so big, that message has been checked whether spam or not.

After that, the relationship has been checked whether is there any relationship between label and length. According to the chart the high message length indicates that it is tend to be spam.

After this part, Text Transformation part comes, it is basically data cleaning/removing of the unimportant data.

In the cleantext function part, if the message has other than the alphabet, ıt is replaced with space and then all the characters turn the lowercase form then they split into words. After that for the list comprehension snowballstemmer is used and then it is returning splitted words together again and the cleantext function is being applied for all the messages. After that messages are compared with before and after.

For the creating Wordcloud, firstly spam and ham messages are seperated with addign filter. After that 2 variables have been created for which words are gonna come in spam messages and ham messages.

Then words blank lists are gonna be created what are the words that are the coming in the spam and ham messages.

ExtractSpamWords & extractHamWords are the two fuctions that extracts the words and then spam&ham words have been referenced in order to use in this function and then list comprehension created with tokenize technique, its basicly converts paragraphs into sentences and sentences to the words then they are conctenated. these functions applied all the spam messages and ham messages seperately.

Wordcloud is created with giving a name as spam. wordcloud for spam messages, and it is also done for ham messages too with giving a name as ham. Then to be able to show which words are mostly used in a sentence, pyplot is used for showing the image.

For the encoding category if else statement is used. If it is spam it returns 1 or returns 0.

Machines only understand from the numbers so the texts should be converted into numerical forms so in this part the text vectorization nlp comes. With the counter vectorizer , it counts the number of messages and the number of unique words. In the next step all the unique words in the sentences are shown with index numbers in the array structure . To be able to use array it must be converted into dataframe structure in order to do that pandas used. Then the new variable ‘len’ which is refers to length is created.

For the output, y variable is created with using label.Firstly Test and Train data are splitted and then 80% for the train data and 20% for the test data used in this model applying part.

Then The Naive Bayes classifier’s MultinominalNB is used for to train the model. Then this model took the name of spam\_detect\_model and once it has trained the predict function can be used.

After that the y\_pred is the predicted value, its found by spam\_detect\_model with predict(x\_test) data. It has predicted so the accuracy can be found from this point.

The dataset which is consists of spam and ham messages are unbalanced, so fbeta\_score is used with the rate of 0.5.

The classification report is also used in here due to unbalanced dataset. F1-score is the harmonic avg of the precision and recall ,and the support is the number of instances of the actual answer found in that class.

From this step model have to be deployed ,and pickle file have to be created and the knowledge of the model has to be created in order to use whole dataset in the web server and the knowledge have to be transformed into pickle file.

Saved\_model is the name of the defined saved model and with the dumps function, objects are stored in a file. Then, the first thing is to check whether saved model is true or not so new model which is modelfrom\_pickle is created and for the knowledge, saved\_model has been used. Then for the modelfrom\_pickle’s prediction, x\_test model is applied.

Finally for the saving pickle file, joblib has imported then again dumps function is used for storing in a file which is pickle, and the used model is that spam\_detect\_model.

The model (learning framework) has been created, for the heroku cloud part, pickle file and app.py which is an instance of the Flask object has to be deployed. App.py will act as the central configuration object for the entire application, and for the surface of the app page, HTML & CSS going to be used.

The final form of the app link: <https://spam-detectionn.herokuapp.com/>

Github Repository link: https://github.com/Erdicem/Spam-Classifier.git

REFERENCES

Expert.ai Team, (May, 2020). What is Machine Learning? A Definition. Retrieved May

15, 2020 from <https://www.expert.ai/blog/machine-learning-definition/>

Lexalytics, (September, 2020). Machine Learning (ML) for Natural Language

Processing (NLP). Retrieved May 15, 2020 from <https://www.lexalytics.com/lexablog/machine-learning-natural-language-processing>

Sharma, R. (July, 2020). Spam classification machine learning | End to End

Deployment Part – 1&2 | Spam Classifier using Python. Retrieved May 15, 2020 from

<https://www.youtube.com/redirect?event=video_description&redir_token=QUFFLUhqbWZxVXZxLS04T1kwN2ctQ2pYOFAzaWlCcFR2UXxBQ3Jtc0trbDc3RDVCYUNQWlowNXk0dllOdUdzSlBQemNGMVFWQjVqWFhnUTJaSHBDR0pMdWNQbXp0ellweEZReXM5bGMtZzdDMWhBNmw5M25xaTQyQkNhRlRjZ0FyX0tqQy0xNEFhZmlqQUROMlpXRFl0OHB5TQ&q=https%3A%2F%2Fdrive.google.com%2Fdrive%2Ffolders%2F1Cz7kAIyx-1JSIdM5fy_hiE75LPR66SjF>