

AI PROJECT

Email Spam Classification

Abstract

Detect Email Spam or Ham from input text.

Team Members:

Mariem Naeem Senada	2022170419
Simon Hanna Reyad	2022170201
Amira Tharwat Hanna	2022170076
Mohamed Osama Ahmed	20201700664
Beshoy Akram Alfy	2022170102
Youssef Ashraf Abdo	2022170511

Pre Processing :

First (Data Cleaning) :

Read file

Know size of data

Know Data frame Features

Data frame Describe

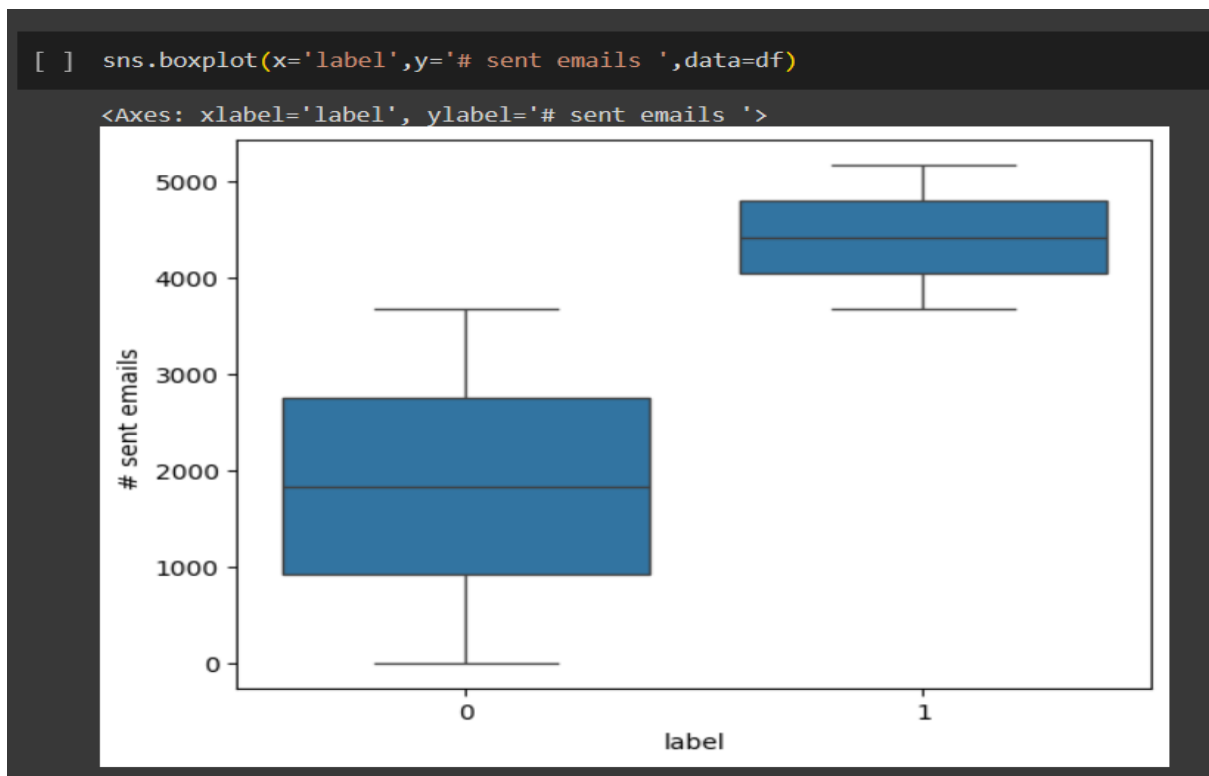
Check nulls

Check duplicate

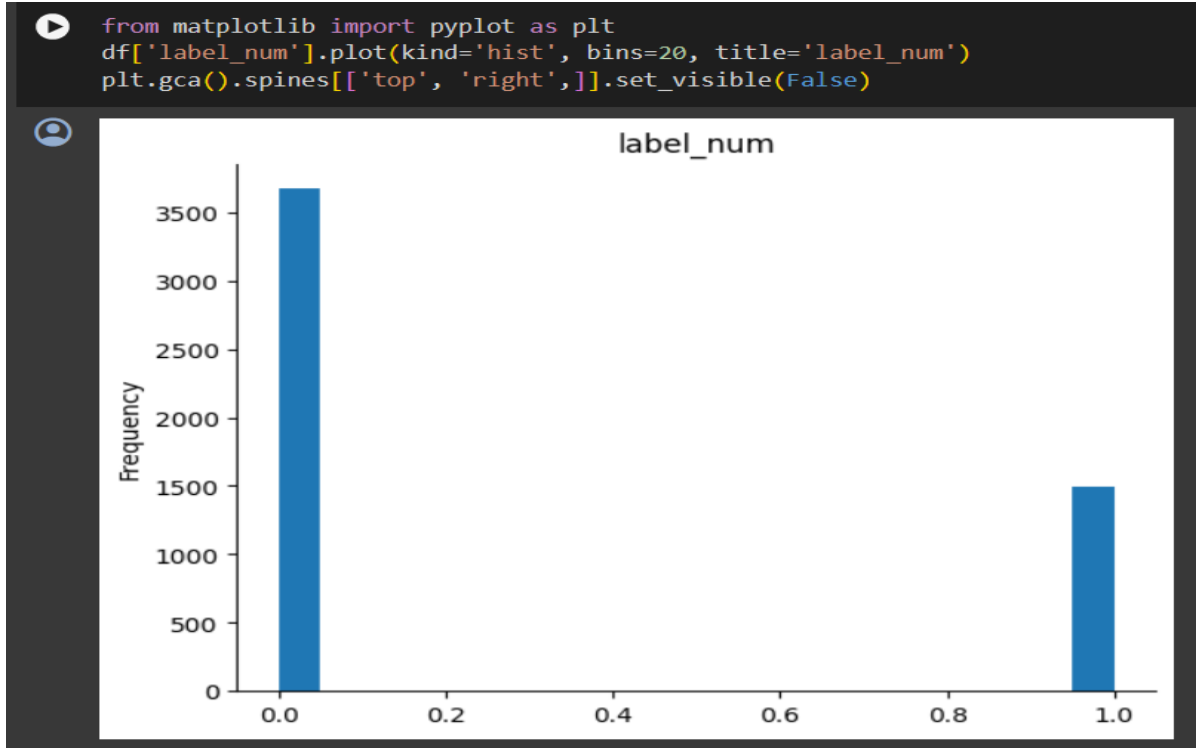
Mapping Data to numeric data (spam =1 , ham =0)

Second (Data representation)

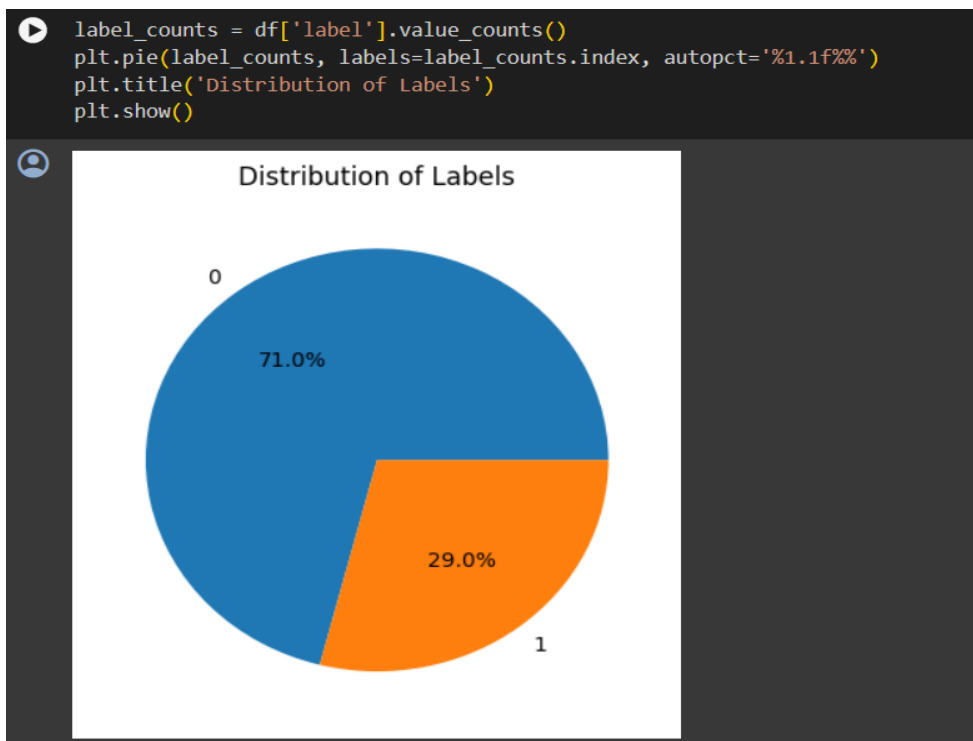
Using boxplot to check outliers of sent emails



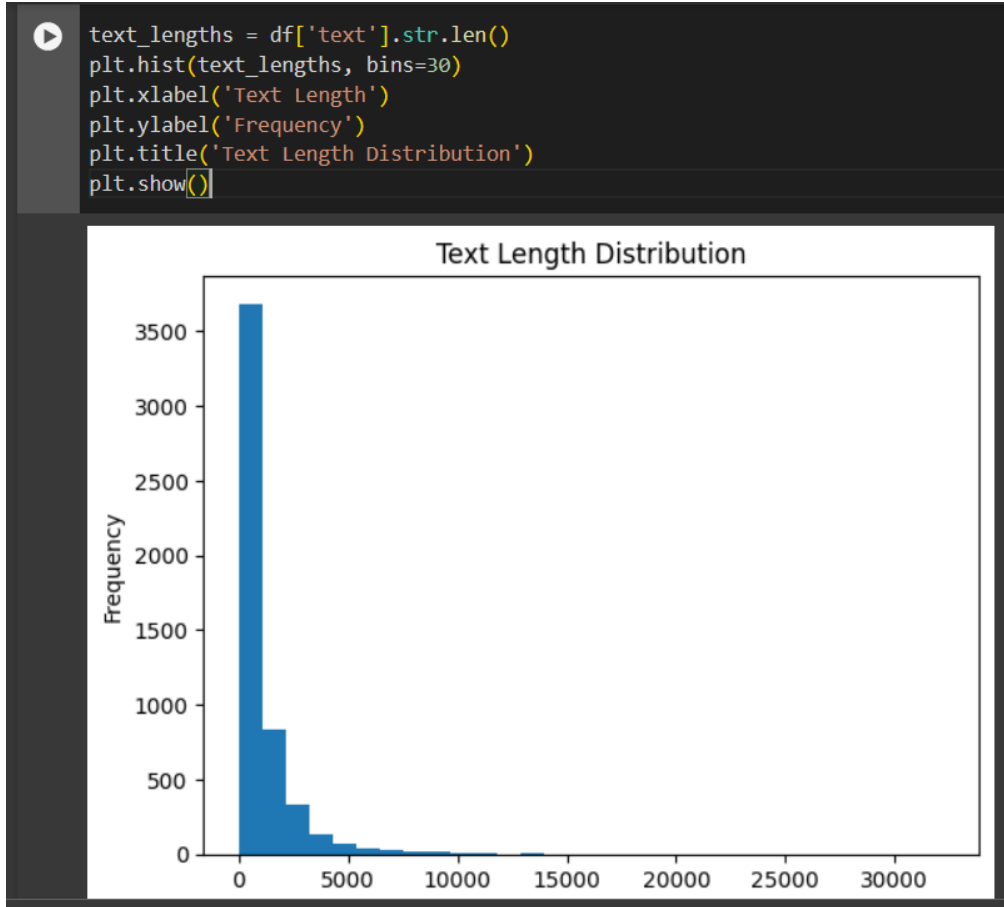
Label Num of data frame (Count spam and Ham)



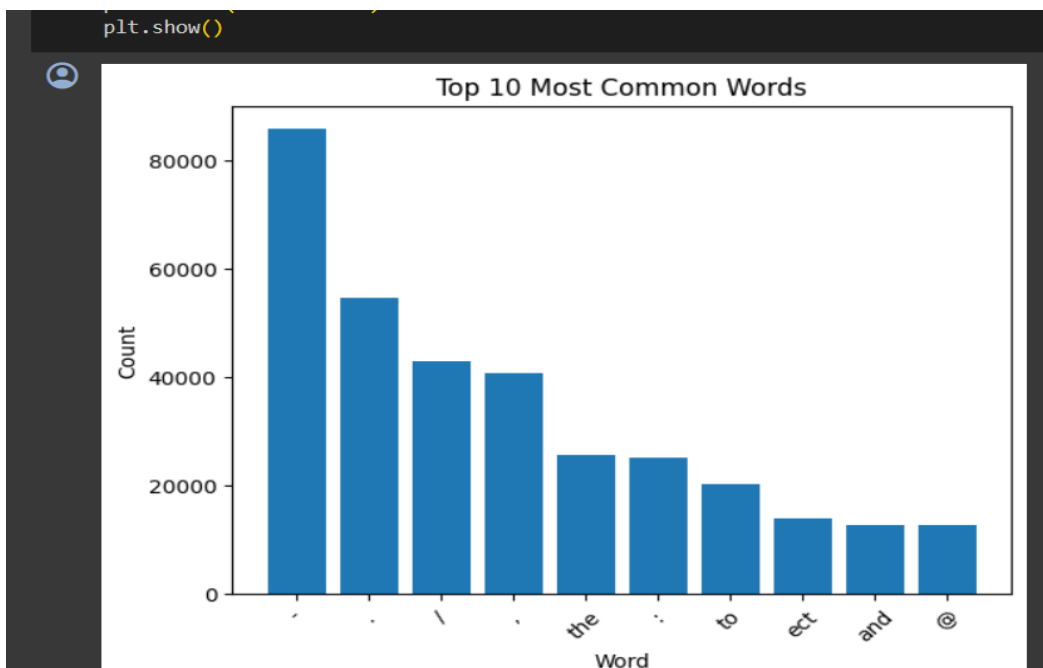
Using pie chart (Percentage of Ham and Spam)



Using Histogram (Frequency of text length Distribution)



Most Common words in text



Models

Using Logistic Regression to classify Data (Spam or Ham)

- Hyper Parameters (C=1.0, solver='liblinear')

Using Decision Tree

- Hyper Parameters (criterion='entropy')

Using SVM

- Hyper Parameters (kernel='rbf', C=1.0)

Using Random Forst

- Hyper Parameters (n_estimators=100, random_state=42)

Using KNN

- Hyper Parameters (n_neighbors=5)

Using Techniques

- vectorize to split text to words and map them into weights
- K best and correlation to specify Feature in algorithm

Conclusion

Project take text input from user and detect it (Spam or Ham) using x-train and y train in the model and using many algorithms (Logistic regression , SVM , KNN, Decision Tree , Random Forst).

Thank you