

***Roll no.-2017647 class roll no.-9***

**Acknowledgement**

First of all, I would like to thank our institution “Graphic Era deemed to be university” for giving us an opportunity to work on such great projects. I am also thankful to our mentor Garima ma’am for guiding me through the project.

**Motivation**

I chose this project because I feel like this project will help me understand more about languages like python, and its different user friendly libraries. The project also helps me understand a little more about how to work with a dataset and use it to analyse and classify messages.As I really like to work with data, this project was an amazing chance to learn more about data and how they work, here, specifically for spam classifiers.

This project also helps me to understand and deploy a website through stream lit and heroku.

**Table of content**

*1. Data cleaning*

*2. EDA*

*3. Text Pre-processing*

*4. Model building*

*5. Evaluation and Improvement*

*6. Website*

*7. Deploy*

***1.Data cleaning-*** For the given dataset ,I’ll be first checking number of columns.

Then , I’ll be removing extra columns from the dataset to reduce the size of the set .

After that I’ll be replacing the first two columns with target and text .

Target will tell us if the message is spam or ham and text column will contain all the messages.

Then I’ll encode label target to spam as 0 and ham as 1 and number all messages from 0 to n-1(where n is total number of messages.

I’ll also check for duplicates and remove if any duplicates are found.

***2.EDA-Exploratory data analysis***. Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

To do this, I’ll be using the matplot and nltk libraries.

First, I’ll check if the data is balanced or not. Here it is not balanced.

Then , I’ll create 3 more columns for counting number of characters , words and sentences.

I’ll analyse the data using seaborn.

Axis subplot :

1.xlabel='num\_characters', ylabel='Count.

2.xlabel='num\_words', ylabel='Count'.

Pairplot.

heat map.

***3.Data Processing-***the carrying out of operations on data, especially by a computer, to retrieve, transform, or classify information.

* Lower case
* Tokenization
* Removing stop words and punctuation
* Removing special characters
* Stemming

All can be achieved by using define function to transform the text to lower case , tokenize ,remove all stop words and punctuations, special characters and stemming using loop ,String ,nltk porter stem and nltk stopwords.

**def** transform\_text(text):

text **=** text**.**lower()

text **=** nltk**.**word\_tokenize(text)

y **=** []

**for** i **in** text:

**if** i**.**isalnum():

y**.**append(i)

text **=** y[:]

y**.**clear()

**for** i **in** text:

**if** i **not** **in** stopwords**.**words('english') **and** i **not** **in** string**.**punctuation:

y**.**append(i)

text **=** y[:]

y**.**clear()

**for** i **in** text:

y**.**append(ps**.**stem(i))

**return** " "**.**join(y)

then I’ll be visualising the words from most used to least used using the word cloud and bar plot for both spam and ham.

***4.Model Training***-I’ll be using sklearn library to train my data set.

After calculating precision , accuracy and confusion matrix for GaussianNB, MultinomialNB , BernoulliNB.

the methods used are-

* LogisticRegression
* SVC
* MultinomialNB
* DecisionTreeClassifier
* KNeighborsClassifier
* RandomForestClassifier
* AdaBoostClassifier
* BaggingClassifier
* ExtraTreesClassifier
* GradientBoostingClassifier
* XGBClassifier

***5.Evaluation and improvisation-***I’ll use catplot and performance to evaluate our model .

# 1. Change the max\_features parameter of TfIdf

# Voting Classifier- to check which is the best classifier method.

# Applying stacking

For now I’ll stick to best method we get , which is multinominal one , but there are more ways to increase accuracy and precision on which I’ll be working on after this evaluation for sure.

Atlast I’ll pickle out transform function and multinominal classification model.

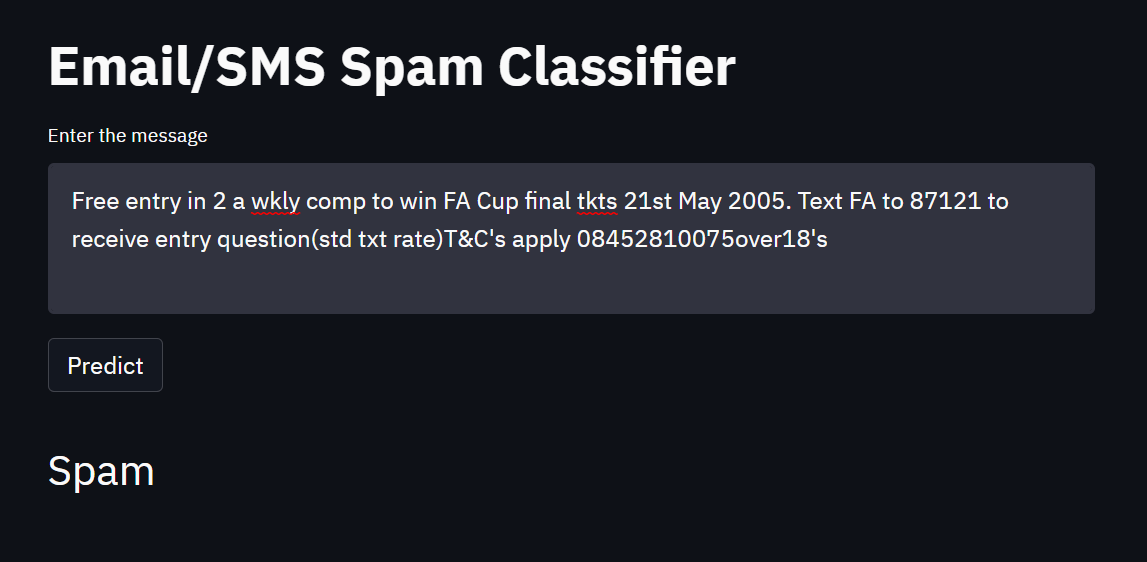
***6. Website-***

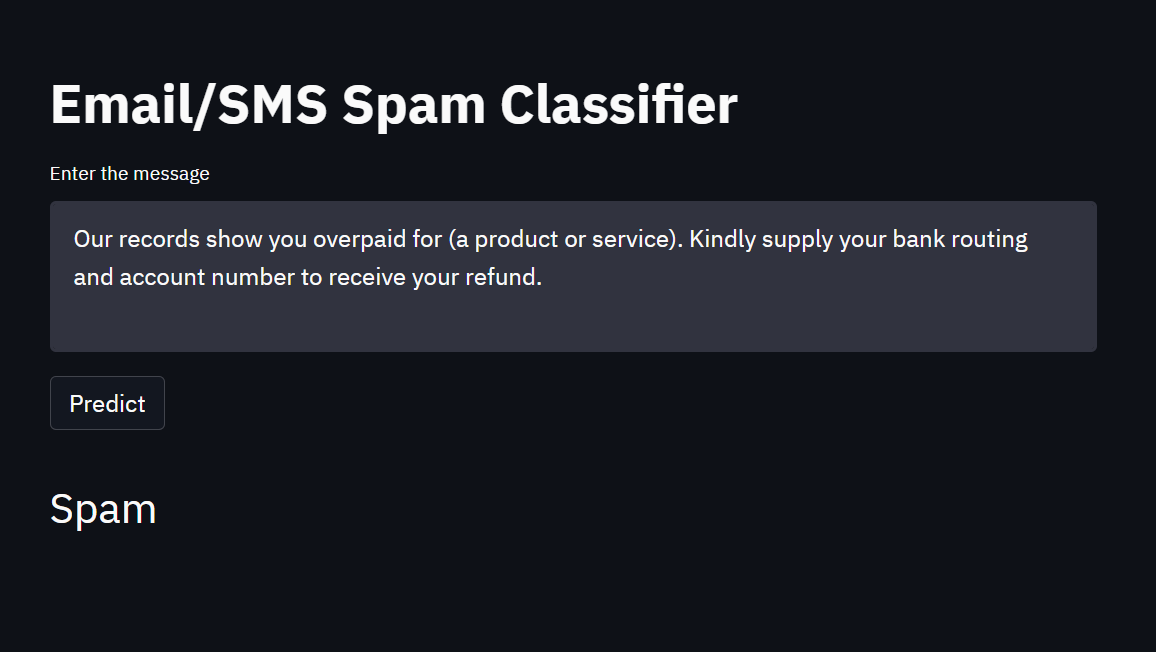
* Preprocess(transform)
* Vectorize
* Predict(using multinominal)
* Display(streamlit)

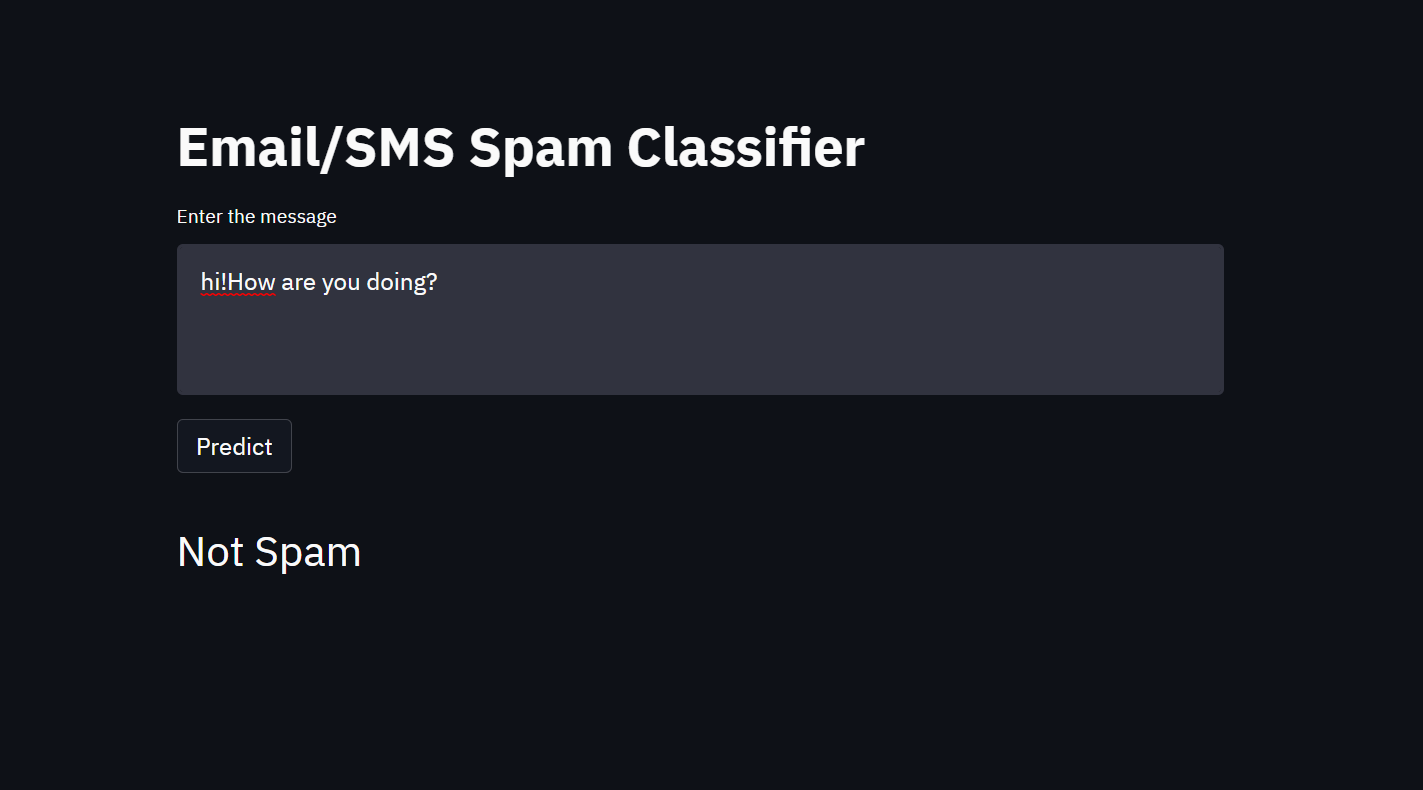
I’ll be using pycharm and streamlit library to make the website.

***7.Deploy-***I’ll use the HEROKU to deploy an app/website for anyone to check if the messages are spam or not.

Examples:







Data used for this classifier is taken from the Kaggle dataset-

https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset