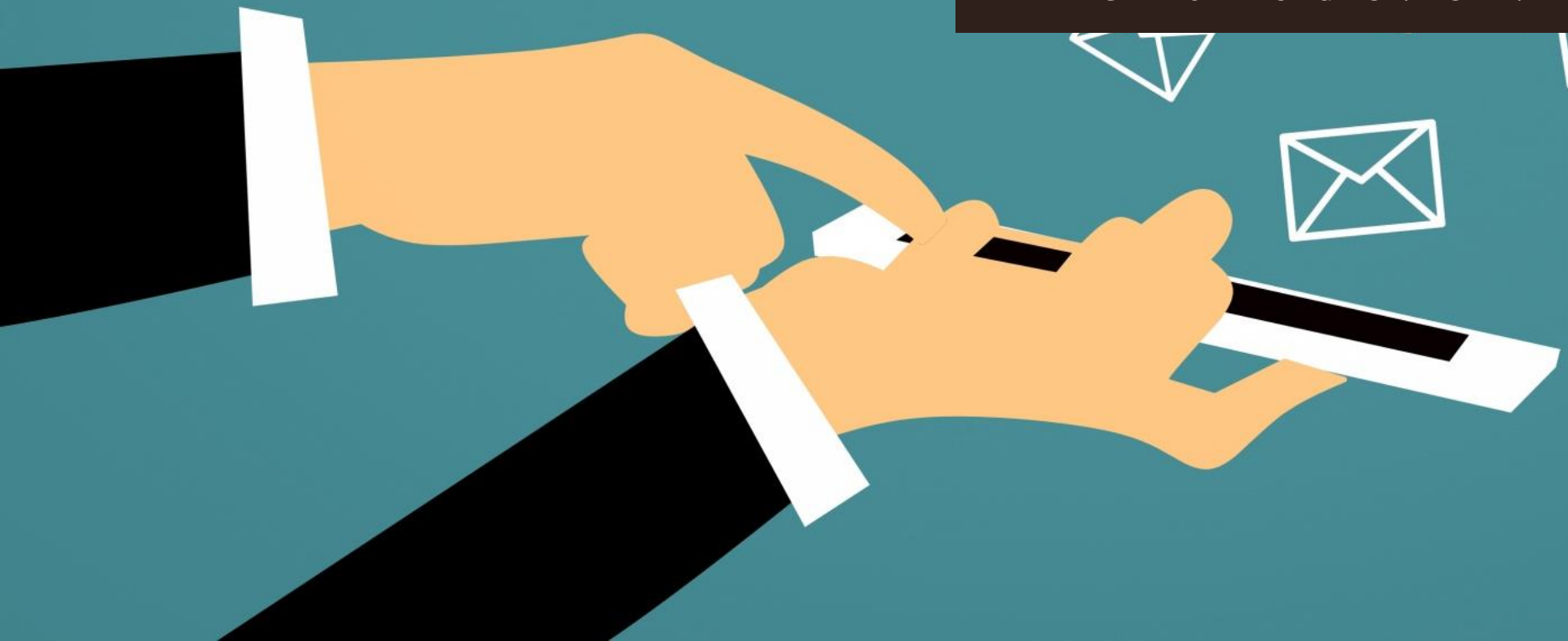
SMS & EMAIL SPAM CLASSIFIER

SUBJECT: DATA 606(CAPSTONE PROJECT)

PROFESSOR: CHAOJIE WANG

PRECENTOR: DURGA SIVA SAI VARMA RUDRARAJU



1.Introduction

The objective of the Email Spam Classifier and SMS Spam Classifier projects is to develop machine learning models and algorithms that can automatically categorize emails and texts as spam (unwanted, unsolicited) or legitimate (non-spam) depending on their content and features. The main objective is to build a platform that classifies out undesirable messages, which will enhance email and messaging for users.

Agenda

1. Introduction
2. Data Description
3. Exploratory Data Analysis (EDA)
4. Model Training
5. Prediction
6. Web Application
7. Conclusion

2.Data Description

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

Data Size: Approximately **503 KB**

Data Description: The "spam.csv" dataset contains SMS messages or emails, particularly focusing on spam classification

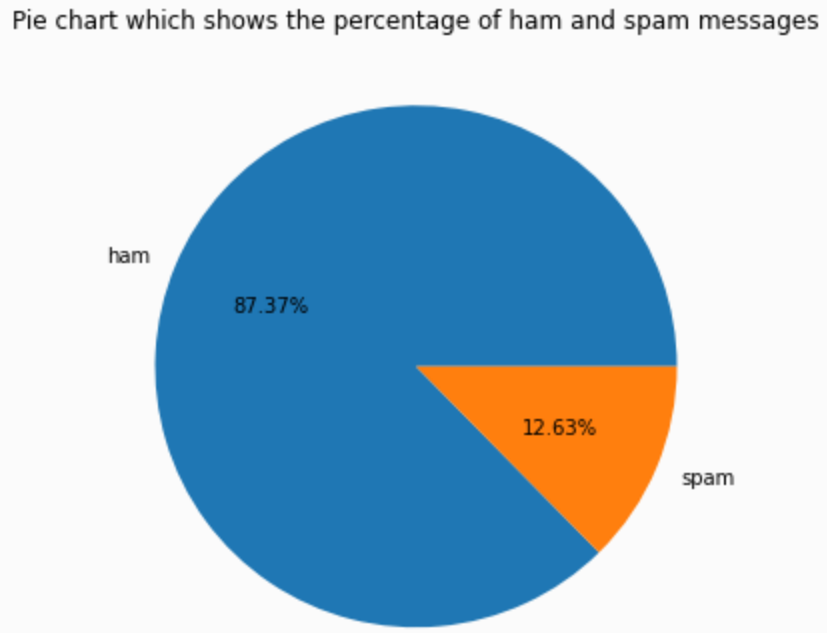
Dataset Size: Number of Rows : **5572** Number of Columns : **2**

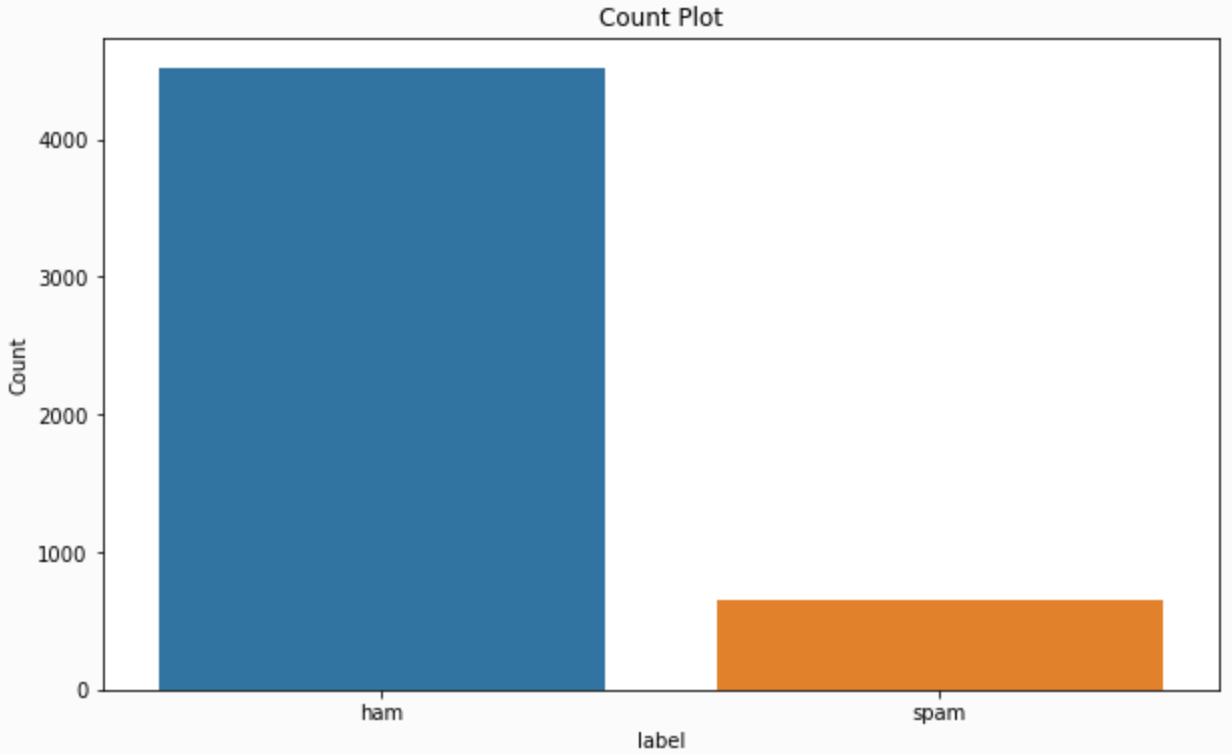
Data Types: v1 : Categorical v2 : Text data

3.EXPLORATORY DATA ANALYSIS

The number of Ham and the number of Spam messages are being counted

- Count plot: The below count plot represents the count vs label graph
- Pie Chart: The pie chart represents the percentage of Ham and Spam messages that are present in the data

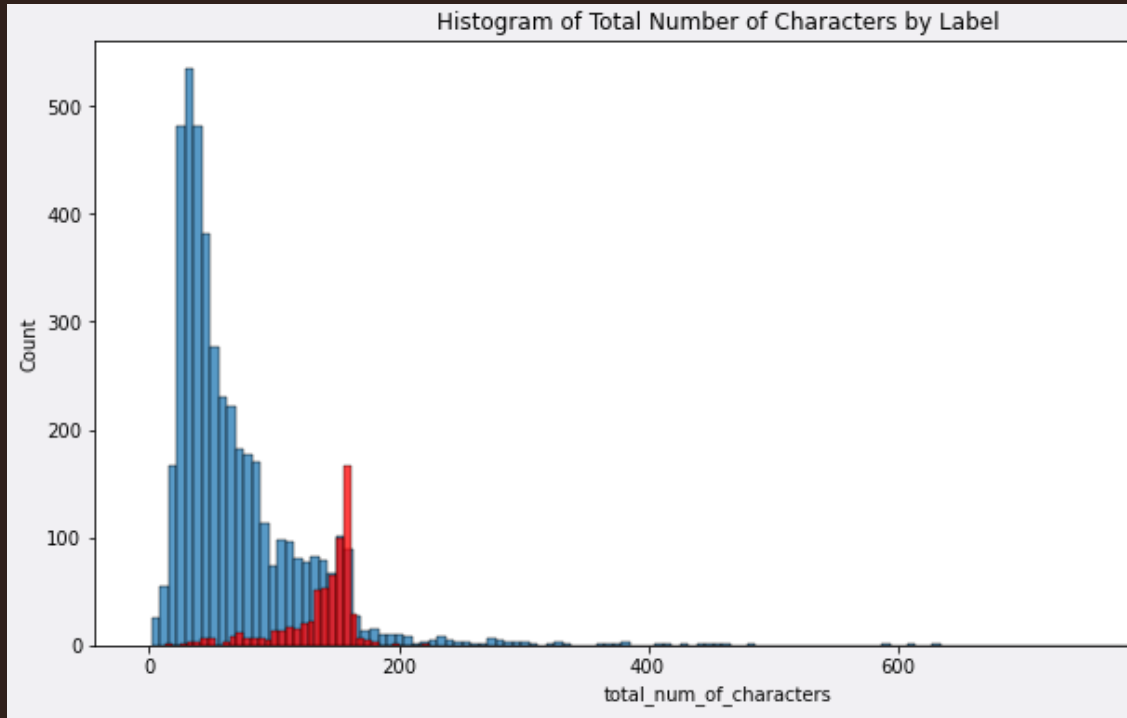




3.EXPLORATORY DATA ANALYSIS

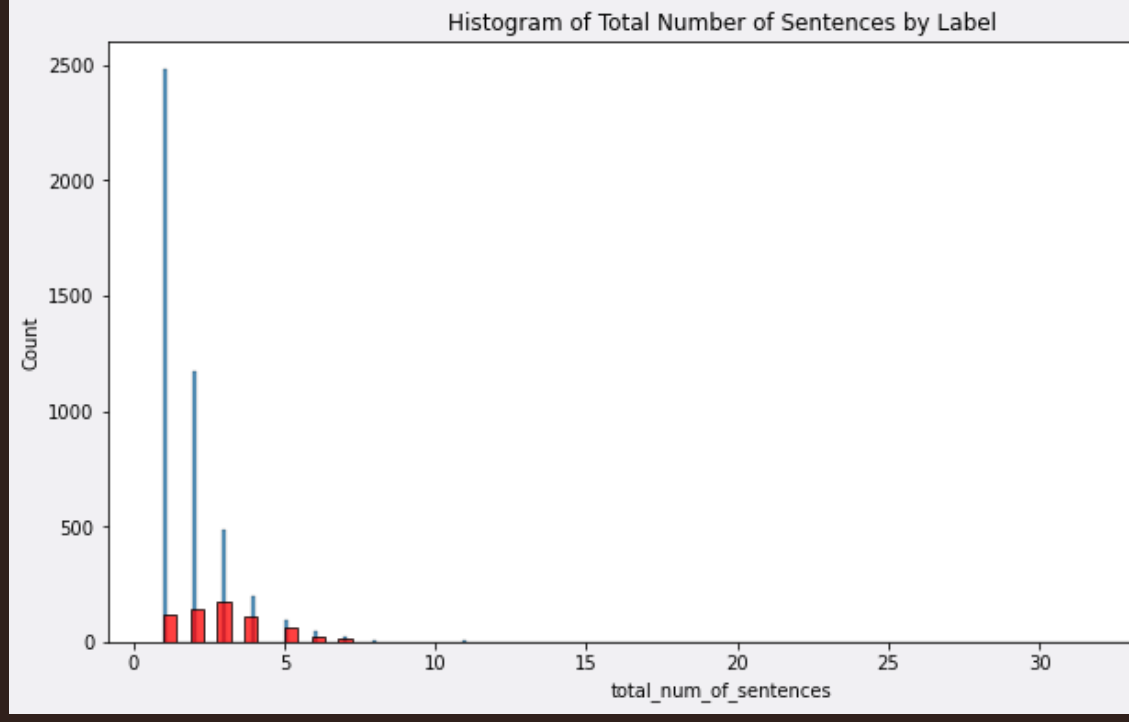
- NLTK stands for Natural Language Toolkit. It is used for working with human language data, particularly for tasks related to natural language processing (NLP). NLTK provides tools, resources, and libraries for a wide range of NLP tasks, including tokenization, stemming, tagging, parsing, semantic reasoning, and more

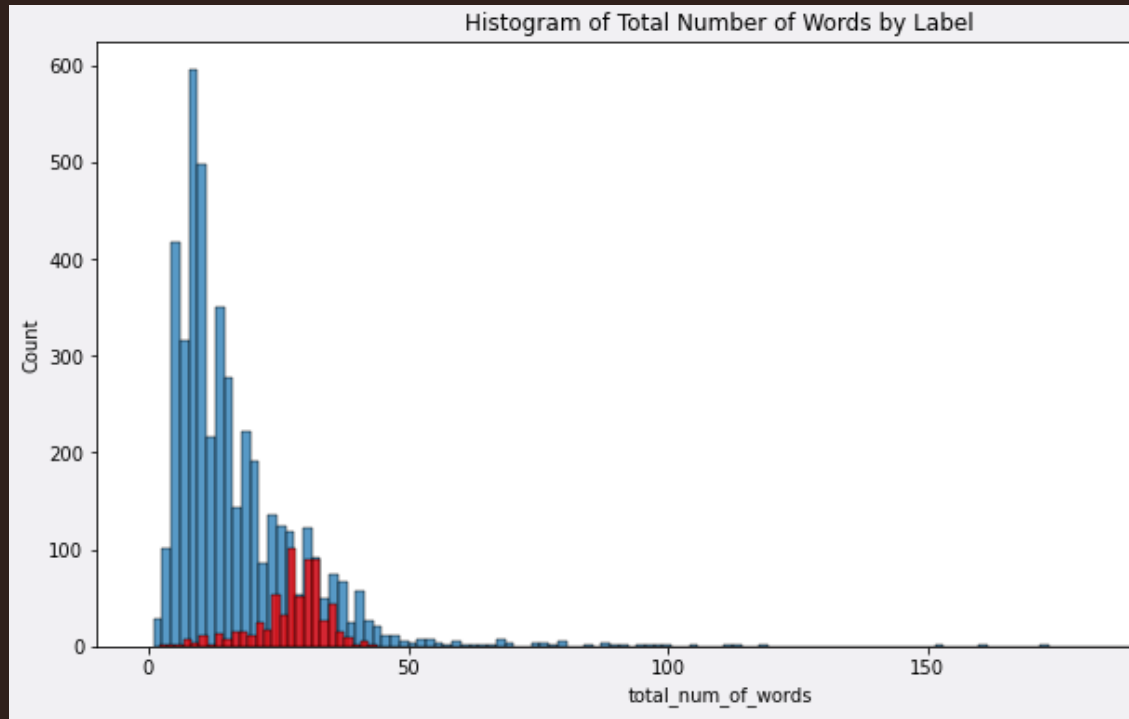
total_num_of_characters	total_num_of_words	total_num_of_sentences	transformed_text
111	23	2	go jurong point avail bugi n great world la e ...
29	8	2	ok lar joke wif u oni
155	37	2	free entri 2 wkli comp win fa cup final tkt 21...
49	13	1	u dun say earli hor u c alreadi say
61	15	1	nah think goe usf live around though
...	...	...	...
161	35	4	2nd time tri 2 contact u pound prize 2 claim e...
37	9	1	b go esplanad fr home
57	15	2	piti mood suggest
125	27	1	guy bitch act like interest buy someth els nex...
26	7	2	rofl true name



3.EXPLORATORY DATA ANALYSIS

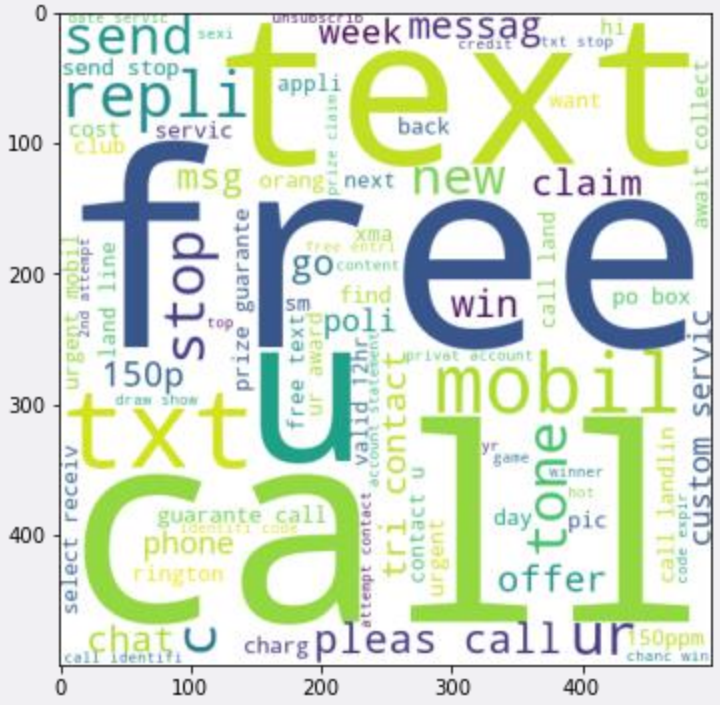
- Plotted the graphs for Histogram of the Total Number of Characters by Label, Histogram of the Total Number of Words by Label, Histogram of Total Number of Sentences by Label and Heatmap of number of characters, number of words, number of sentences columns





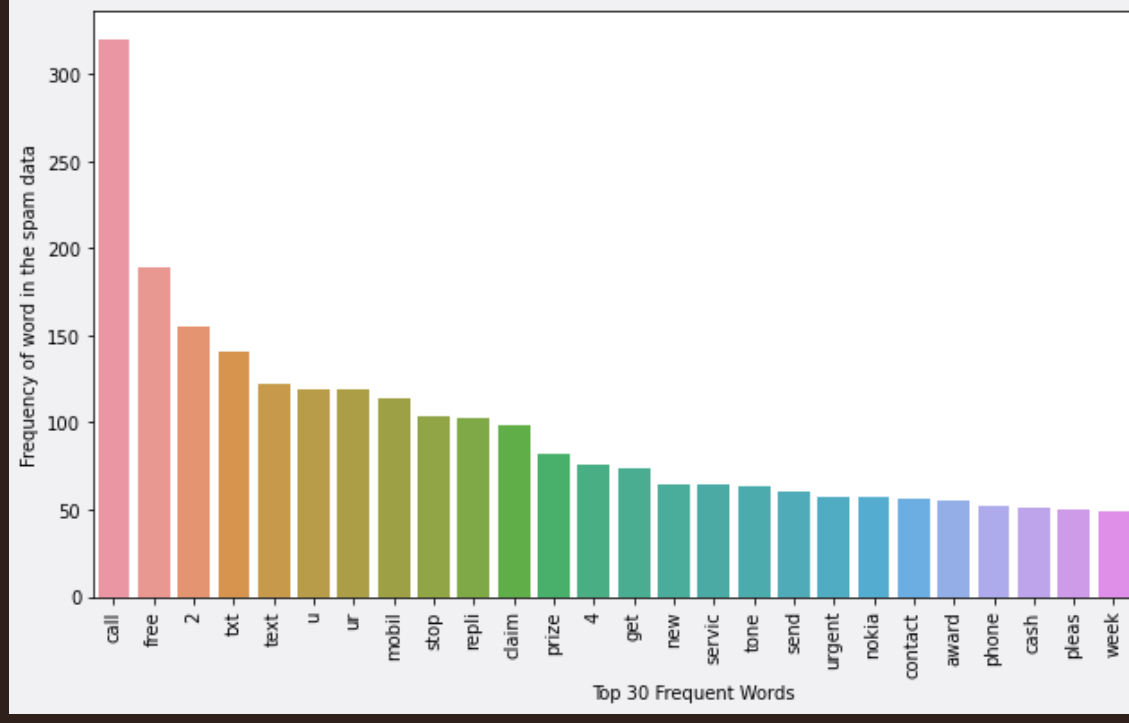
3.EXPLORATORY DATA ANALYSIS

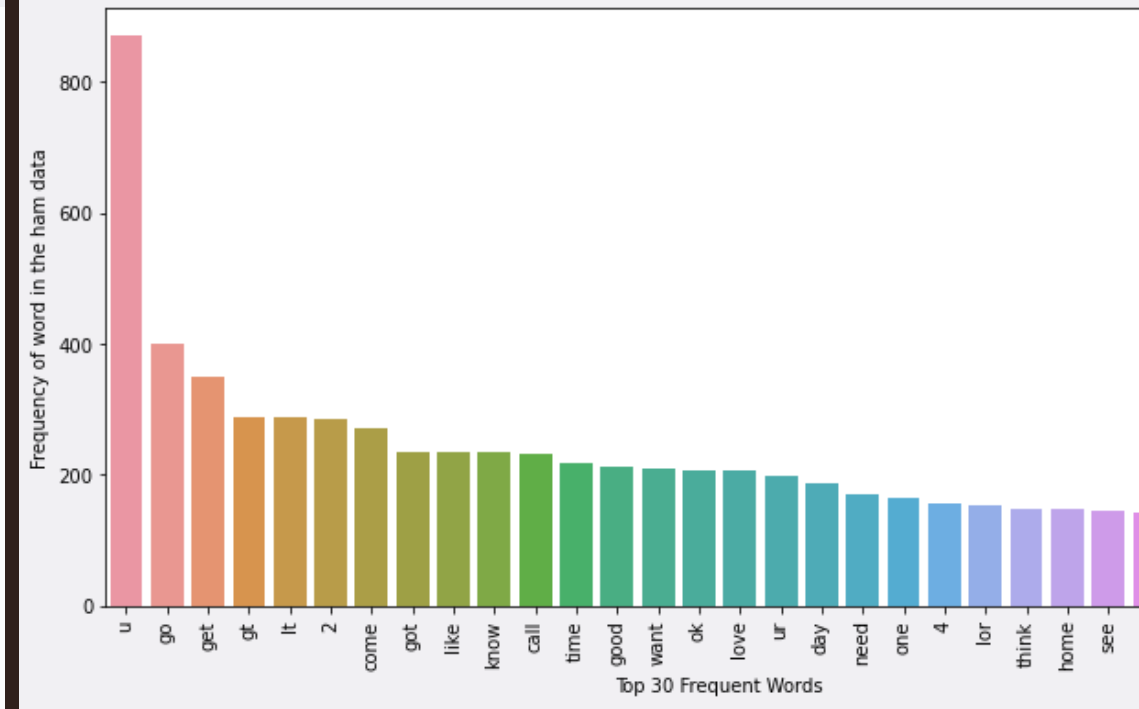
- Displaying the wordcloud is the graphical representation of word frequency that give greater preference to words that appear more frequently in a source text. Using the wordcloud library I am creating two wordclouds one which shows spam words and the other which shows non-spam words.



3.EXPLORATORY DATA ANALYSIS

- Created a list of spam and ham words from the data and found the Total spam words in the data and Total ham words in the data
- Plotted the Bar plots for the common words in the spam & ham text: The bar plot represents the Top 30 Frequent Words V/s Frequency of word in the spam & ham data.





4.MODEL TRAINING

How will you train the models?

Train vs test split: 80% of the data for training the models and 20% for testing

Python packages to be used: scikit-learn for building, training, and evaluating machine learning models.

The development environments: Jupyter Notebook environment with access to crucial tools and resources for data science and machine learning

How will you measure and compare the performance of the models?
: Accuracy, Precision and Confusion_matrix

4.1 Models Used for Machine Learning

Initially I will be using Navie bayes model as we know that for text based dataset Navie bayes works well later other models will also be used.

GaussianNB Scores

- Accuracy Score: 0.86
- Precision Score: 0.5

MultinomialNB Scores

- Accuracy Score: 0.97
- Precision Score: 1.0

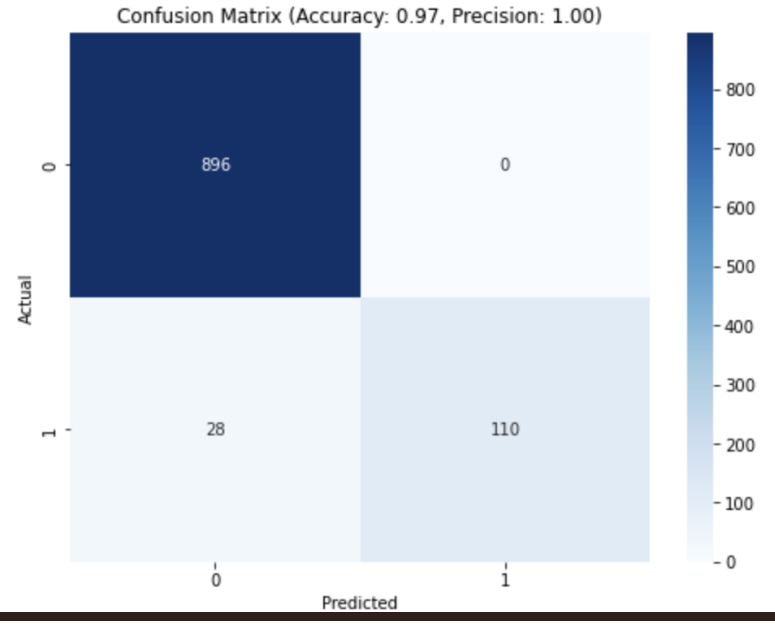
BernoulliNB Scores

- Accuracy Score: 0.98
- Precision Score: 0.99

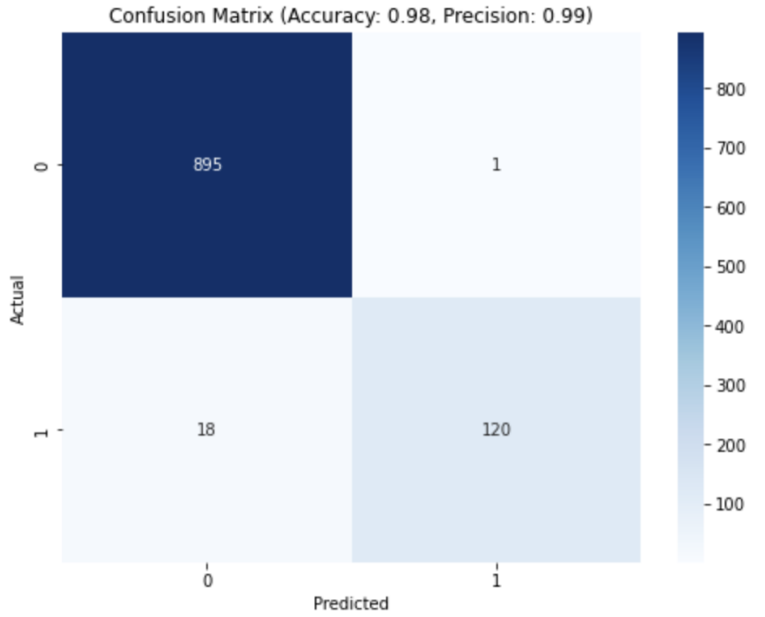
4.1 Models Used for Machine Learning

Confusion matrix for both MultinomialNB and BernoulliNB

MultinomialNB Accuracy Score: 0.9729206963249516
MultinomialNB Precision Score: 1.0



BernoulliNB Accuracy Score: 0.9816247582205029
BernoulliNB Precision Score: 0.9917355371900827



4.1 Models Used for Machine Learning

Here we can see the comparison of the performance of all the models that have been used. From the table we can say that Navie bayes gives the best precision and accuracy.

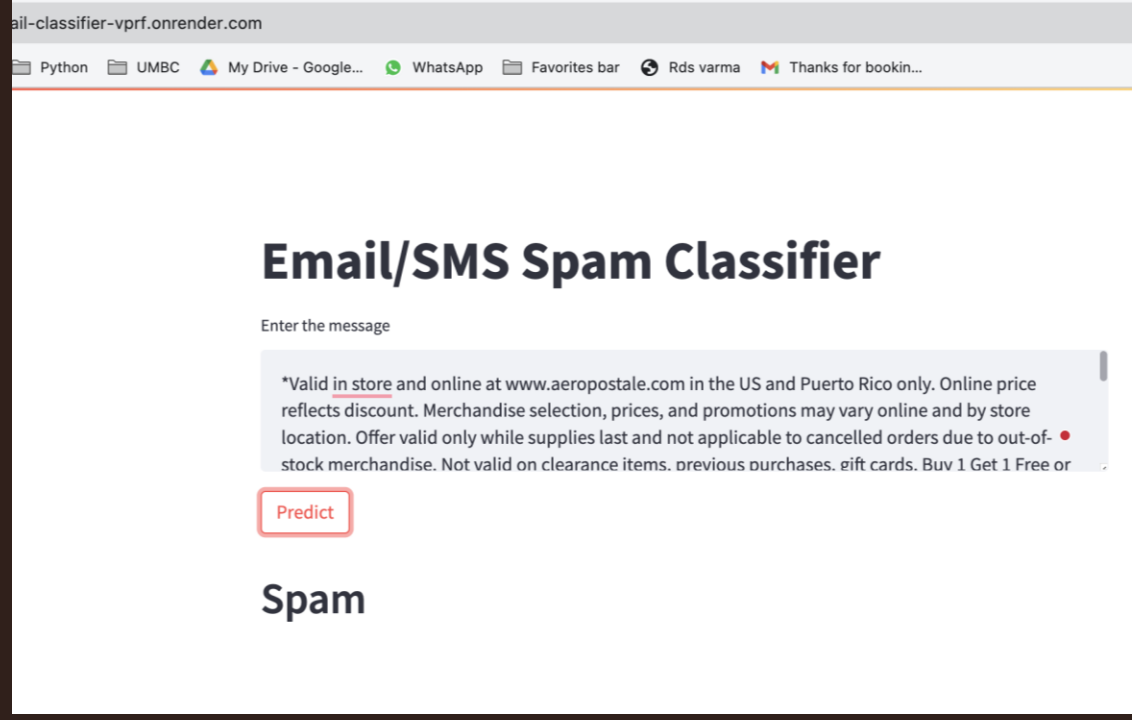
```
performance_df
```

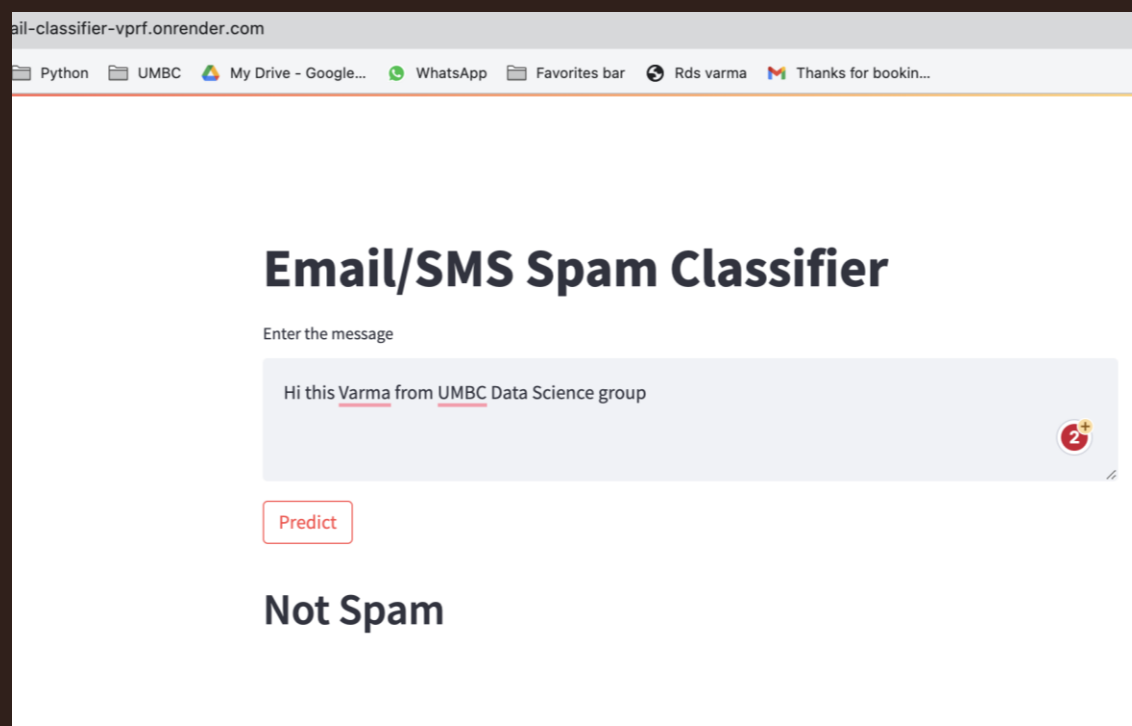
	Algorithm	Accuracy	Precision
1	KN	0.905222	1.000000
2	NB	0.972921	1.000000
6	AdaBoost	0.966151	0.981308
5	RF	0.972921	0.974138
8	ETC	0.975822	0.966942
0	SVC	0.974855	0.966667
4	LR	0.956480	0.951456
10	xgb	0.968085	0.941176
9	GBDT	0.951644	0.931373
7	BgC	0.955513	0.859375
3	DT	0.932302	0.854167

5. WEB APPLICATION

I have created a web application using Streamlit. This is a user friendly interface where a user can just give the text and check if massage they have received is spam or not spam.

https://sms-email-classifier-vprf.onrender.com/





6. CONCLUSION

- Successfully conducted the training and testing phases of the project.
- Employed a machine learning model for accurate prediction of whether a message is spam or not
- Implemented a user-friendly interface for seamless interaction.
- Users can input a message into the text box to instantly check its spam classification.


THANK YOU !!!
ANY QUESTIONS ?