**A MINI PROJECT ON**



**SMS/EMAIL CLASSIFIER**



**Submitted in partial fulfilment of the requirement for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by:**

**Ishita Bhatt** **20022638**

***Under the Guidance of***

**Dr. Surendra Kumar Shukla**



**Department of Computer Science and Engineering**

**Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand**

**September-2023**



**CANDIDATE’S DECLARATION**

I/We hereby certify that the work which is being presented in the Synopsis entitled **“SMS/Email Classifier”** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineeringin the Department of Computer Science and Engineering of the Graphic Era (Deemed to be University), Dehradun shall be carried out by the undersigned under the supervision of **Dr.** **Surendra Kumar Shukla, Professor**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Ishita Bhatt 20022638

The above mentioned students shall be working under the supervision of the undersigned on the **“SMS/Email Classifier”**

**Supervisor**

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction and Problem Statement | 4 |
| Chapter 2 | About The Dataset | 5-6 |
| Chapter 3 | Methodology | 7-9 |
| Chapter 4 | Result and Conclusion | 10-11 |

**INTRODUCTION AND PROBLEM STATEMENT**

**INTRODUCTION:**

In today's digital age, Short Message Service (SMS) communication is a ubiquitous and essential means of exchanging information. People use SMS for various purposes, including personal communication, business transactions, and service notifications. As a result, the volume of SMS messages being sent and received has grown tremendously, leading to the need for efficient management and organization of these messages.

An SMS classifier is a critical tool in addressing this challenge. It is a machine learning-based system designed to automatically categorize or classify incoming SMS messages into predefined categories or labels. The goal is to streamline the handling of SMS messages, making it easier for users to prioritize and respond to messages promptly.

**Problem Statement:**

The problem addressed by an SMS classifier revolves around the overwhelming volume and diversity of SMS messages that individuals and organizations receive on a daily basis.

The problem at hand involves the development of an SMS classifier to manage the increasing volume of SMS messages effectively. Key challenges include accurate message categorization, spam detection, prioritization, customization, and ensuring user data privacy and security. The goal is to enhance the overall SMS experience and streamline message management.

**ABOUT THE DATASET**

Context:

The SMS Spam Collection is a set of SMS tagged messages that have been collected for SMS Spam research. It contains one set of SMS messages in English of 5,574 messages, tagged acording being ham (legitimate) or spam.

Content:

The files contain one message per line. Each line is composed by two columns: v1 contains the label (ham or spam) and v2 contains the raw text.

This corpus has been collected from free or free for research sources at the Internet:

-> A collection of 425 SMS spam messages was manually extracted from the Grumbletext Web site. This is a UK forum in which cell phone users make public claims about SMS spam messages, most of them without reporting the very spam message received. The identification of the text of spam messages in the claims is a very hard and time-consuming task, and it involved carefully scanning hundreds of web pages. The Grumbletext Web site is: [Web Link].

-> A subset of 3,375 SMS randomly chosen ham messages of the NUS SMS Corpus (NSC), which is a dataset of about 10,000 legitimate messages collected for research at the Department of Computer Science at the National University of Singapore. The messages largely originate from Singaporeans and mostly from students attending the University. These messages were collected from volunteers who were made aware that their contributions were going to be made publicly available. The NUS SMS Corpus is avalaible at: [Web Link].

-> A list of 450 SMS ham messages collected from Caroline Tag's PhD Thesis available at [Web Link].

-> Finally, we have incorporated the SMS Spam Corpus v.0.1 Big. It has 1,002 SMS ham messages and 322 spam messages and it is public available at: [Web Link]. This corpus has been used in the following academic researches:

**METHODOLOGY**

**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.

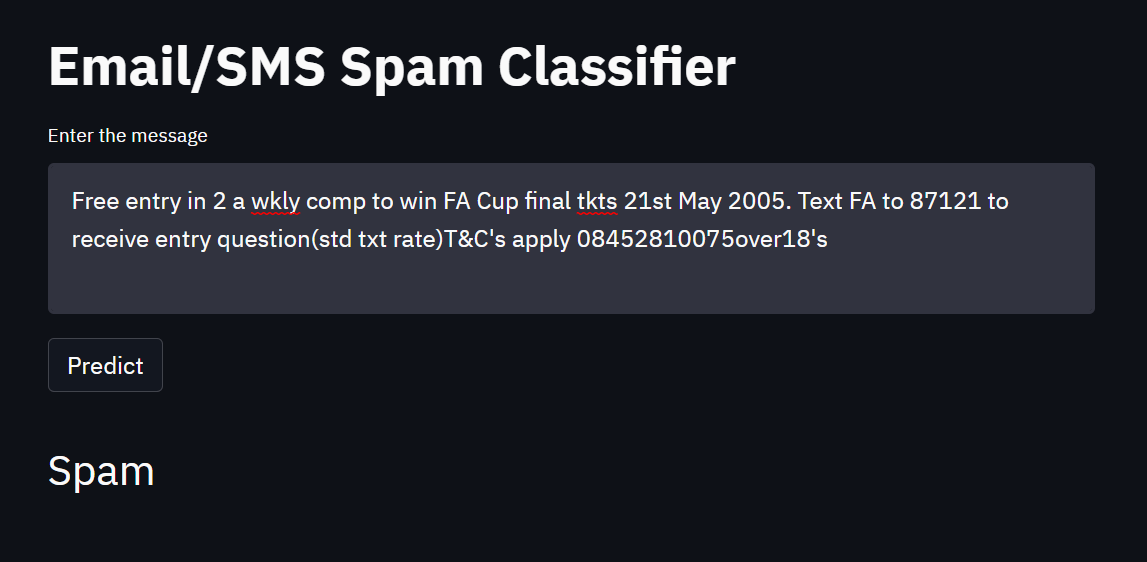
***6. Website-***

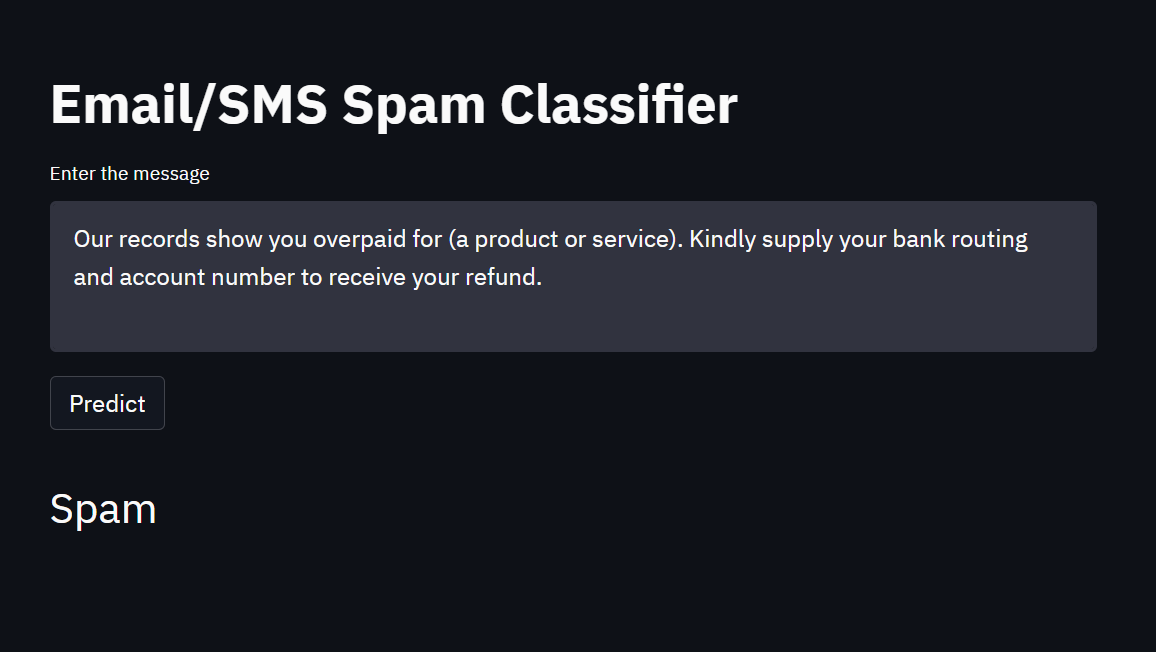
* Preprocess(transform)
* Vectorize
* Predict(using multinominal)
* Display(streamlit)

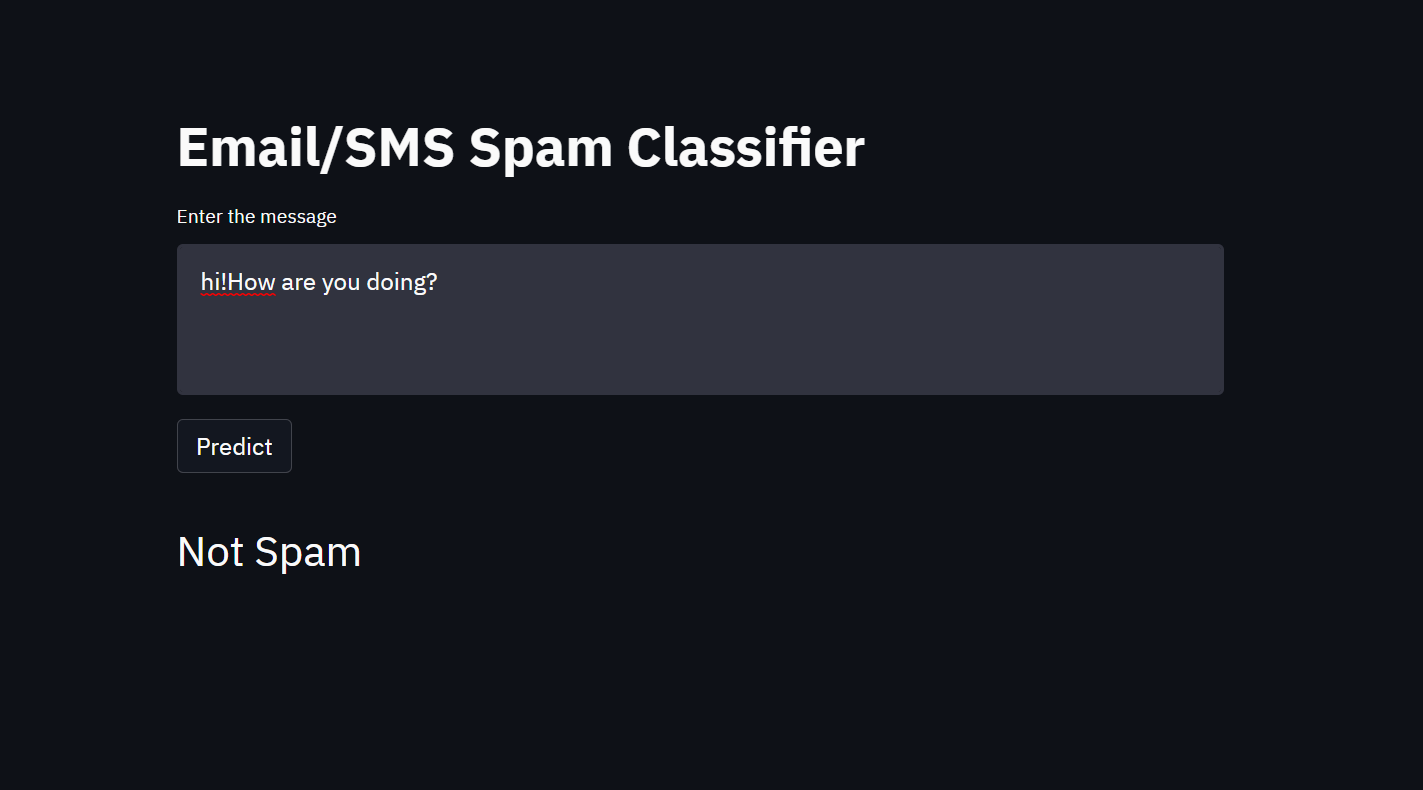
I’ll be using pycharm and streamlit library to make the website.

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

**RESULT AND CONCLUSION**







As we can see, our model works well and is giving accurate predictions.

In conclusion, the development and implementation of an SMS classifier offer significant benefits for users and organizations grappling with the ever-expanding volume of Short Message Service (SMS) messages.

SMS classifiers represent a valuable tool in the modern world of SMS communication. They not only tackle the challenges associated with managing large volumes of messages but also contribute to improved efficiency, security, and overall user satisfaction. As SMS continues to be a crucial means of communication, the development and deployment of effective SMS classifiers are poised to become increasingly important