A Project-I Report

On

EMAIL SPAM CLASSIFICATION USING MACHINE LEARNING

Submitted to

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY ANANTAPUR, ANANTHAPURAMU

In Partial Fulfillment of the Requirements for the Award of the Degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE & ENGINEERING

Submitted By

SANDESH POKHREL	**-**	**(19691A05J0)**
NABIN SHAHI	**-**	**(19691A05J6)**
YOGESH.C	**-**	**(19691A05I5)**
SREEDEVI.H	**-**	**(19691A05F5)**

Under the Guidance of

Dr.D.Jagadeesan,M.Tech., Ph.D.

Professor

Department of Computer Science & Engineering

MADANAPALLE INSTITUTE OF TECHNOLGY & SCIENCE

(UGC – AUTONOMOUS)

(Affiliated to JNTUA, Ananthapuramu)

Accredited by NBA, Approved by AICTE, New Delhi)

AN ISO 9001:2008 Certified Institution

P. B. No: 14, Angallu, Madanapalle – 517325

2019-2023

✆ : 08571 - 280255, 280590
MADANAPALLE INSTITUTE OF TECHNOLOGY & SCIENCE
An Engineering College Sponsored by :
RATAKONDA RANGA REDDY EDUCATIONAL ACADEMY
Approved by AICTE, New Delhi and Affiliated to JNTUA, Anantapur
P.B. No. 14, Angallu, MADANAPALLE - 517 325. Chittoor Dist. A.P.
www.mits.edu

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that the project work entitled **"EMAIL SPAM CLASSIFICATION USING MACHINE LEARNING "** is a bonafide work carried out by

SANDESH POKHREL	**-**	**(19691A05J0)**
NABIN SHAHI	**-**	**(19691A05J6)**
YOGESH.C	**-**	**(19691A05I5)**
SREEDEVI.H	**-**	**(19691A05F5)**

Submitted in partial fulfillment of the requirements for the award of degree **Bachelor of Technology** in the stream of **Computer Science & Engineering** in **Madanapalle Institute of Technology and Science, Madanapalle,** affiliated to **Jawaharlal Nehru Technological University Anantapur, Ananthapuramu** during the academic year 2022-2023

Guide
Dr. D.Jagadeesan,M.Tech., Ph.D
Professor
Department of CSE

Head of the Department
Dr. R. Kalpana, M.E., Ph.D
Professor and Head,
Department of CSE

Submitted for the University examination held on:

Internal Examiner
Date:

External Examiner
Date:

ACKNOWLEDGEMENT

We sincerely thank the management of **Madanapalle Institute of Technology and Science** for providing excellent infrastructure and lab facilities that helped me to complete this project.

We sincerely thank **Dr. C. Yuvaraj, M.E., Ph.D**., **Principal** for guiding and providing facilities for the successful completion of our project at **Madanapalle Institute of Technology and Science,** Madanapalle.

We express our deep sense of gratitude to **Dr. R. Kalpana, M.E., Ph.D., Professor and Head of the Department of CSE** for his continuous support in making necessary arrangements for the successful completion of the Project.

We express our deep gratitude to my guide **Dr.D.Jagadeesan,M.Tech., Ph.D.**, **Professor, Department of CSE** for his guidance and encouragement that helped us to complete this project.

We express my deep sense gratitude to **Dr. K P Manikandan, Ph.D** , **Project Coordinator and Mrs. M Bommy, M.E., Project Co-Coordinator** for their valuable guidance and encouragement that helped us to complete this project.

We also wish to place on record my gratefulness to other **Faculty of CSE Department** and also to our friends and our parents for their help and cooperation during our project work.

SANDESH POKHREL
NABIN SHAHI
YOGESH.C
SREEDEVI.H

DECLARATION

We hereby declare that the results embodied in this project **" EMAIL SPAM CLASSIFICATION USING MACHINE LEARNING"** by us under the guidance of **Dr.D.Jagadeesan,M.Tech, Ph.D**, **Professor, Dept. of CSE** in partial fulfillment of the award of **Bachelor of Technology** in **Computer Science & Engineering** from **Jawaharlal Nehru Technological University Anantapur, Ananthapuramu** and we have not submitted the same to any other University/institute for award of any other degree.

PROJECT ASSOCIATES

SANDESH POKHREL
NABIN SHAHI
YOGESH.C
SREEDEVI.H

Date :
Place :

I certify that above statement made by the students is correct to the best of my knowledge.

Date : **Guide**

INDEX

List of Figures

ABSTRACT

Email is the most used source of official communication method for business purposes. The usage of the email continuously increases despite of other methods of communications. Automated management of emails is important in the today's context as the volume of emails grows day by day. Out of the total emails, more than 55 percent is identified as spam. This shows that these spams consume email user time and resources generating no useful output. The spammers use developed and creative methods in order to fulfil their criminal activities using spam emails, Therefore, it is vital to understand different spam email classification techniques and their mechanism. This paper mainly focuses on the spam classification approached using machine learning algorithms. Furthermore, this study provides a comprehensive analysis and review of research done on different machine learning techniques and email features used in different Machine Learning approaches. Also provides future research directions and the challenges in the spam classification field that can be useful for future researchers.

CHAPTER-1

INTRODUCTION

1.1 MOTIVATION

Email has become one of the most important forms of communication. In 2014, there are estimated to be 4.1 billion email accounts worldwide, and about 196 billion emails are sent each day worldwide . Spam is one of the major threats posed to email users. In 2013 69.6% of all emails flows were spam . Links in spam emails may lead to users to website with malware or phishing schemes ,which can access and disrupt the receiver's computer system. These sites can also gather sensitive information from. Additionally, spam costs businesses around $2000 per employee per year due to decreased productivity. Therefore, an effective spam filtering technology is a significant contribution to the sustainability of the cyberspace and to our society.

Current spam techniques could be paired with content - based spam filtering methods to increase effectiveness . Content-based methods analyze the content of email to determine if the email is spam . The goal of our project was to analyze machine learning algorithms and determine their effectiveness as content-based spam filters.

1.2 OBJECTIVE OF THE PROJECT

- The main objective of this project is to classify the mail content is spam or not.
- Email spam classification is aimed at reducing to the barest minimum the volume of unsolicited emails.
- To give knowledge to the user about the fake e-mails and relevant e-mails .

1.3 LIMITATIONS OF PROJECT

- Block mail from known spam resources.
- Readily available pools of lists.
- It is effective and easy to implement.
- It reduces error rates as legitimate e-mail would not be blocked even if the ISP from which is originated , is on a real time block list.
- The presence of a single token should not cause the e-mail to be classified as spam.
- Blocks known spam.
- It is very effective and is also adaptive,so hard to fool.

1.4 ORGANISATION OF DOCUMENTATION

Feasibility Study

All systems are feasible if they are given unlimited resources and infinite time. There are aspects in the feasibility study portion of the preliminary investigation:

- Technical Feasibility
- Operation Feasibility
- Economic Feasibility

Technical Feasibility

The technical requirement for the system is economic and it does not use any other hardware and software.

Operation Feasibility

The operational feasibility includes User friendly, reliability, security, portability, availability and maintainability of the software used in the project.

Economic Feasibility

Analysis of a project costs and revenue in an effort to determine whether or not it is logical and possible to complete.

CHAPTER-2

LITERATURE SURVEY

2.1 INTRODUCTION

Nowadays,emails are used in almost every field, from business to education.Emails have two subcategories i.e. ham and spam. Email spam, also called junk emails or unwanted emails is a type of email that can be used to harm any user by wasting his/her time, computing resources, and stealing valuable information.In Practice receiving spam email decrease the efficiency of the user and it is quite

annoying to the user.Usually,such emails are forwarded for profit making purposes.Spam classification identification has the major objective to inform users about fake emails/or relevant emails.

2.2 REVIEW OF PREVIOUS PAPER

- The Internet has become an inseparable part of human lives,where more than four and half billion Internet users find it a convenient to use it for their facilitation. Moreover emails are considered as a reliable form of communication by the Internet Users [1].
- Over the decades, email services have been envolved into a powerful tool for the exchange of different kind of information. The increased use of e-mail also entails more spam attacks for the Internet users. Spam can be sent from anywhere on the planet from users having deceptive intentions that has acess to the internet. Spam are unsolicited and unwanted emails sent to receipants who do not want or need them. The emails spam hava fake content with mostly links for phising attacks and other threads, and these emails are sentin bulk to many recipents[2].
- The intention behind them is to steal users' personal information and then use them against their will to gain materialistic benefits[3].
- These emails either contain malicious content or have URLs that lead to malicious content .Such emails are also sometimes referred to as phising emails.Despite the advancementof spam filtering applications and services, there is no definitive way to distinguish between legitimate and malicious emails because of ever-changing content of such emails. Spams have been sent for over three or four decades now, and with the availability of various antispam services,even today,nonexpert end-users get trapped into such hideous pitfall [4].

- Precison ,recall and f-measure and considered key evaluating measures to compare Naive Bayes and SVM, while the evaluations parameters,i.e., Model Loss and ROC-AUC, are calculated for deep learning models such as CNN and LSTM. Finally, a comparision is made between all models for the best accuracy and values for evaluation parameters obtained by DL and ML models [5].
- The authors in [6] gathered 1463 tweets written in Roman Urdu and categorized 1038 of them as ham and 425 of them as spam. On the data, they used discriminative multinomial Naive Bayes techniques. The got 95.12% with DMNBTex and 95.12% with NB. The techniques were used in numerical sequence of words that did not take into account domain or linguistic details.Linguistic techniques, such as those that take into account the contextual characteristics of important terms in Roman Urdu litreature, are expected to improve classsification result.

2.3 EXISTING SYSTEM

- Machine Learning based Spam E-Mail Detection had done by using J48 algorithms.
- It had an average accuracy of 87.5%.
- It has less accuracy when completed to the proposed system.

2.4 DISADVANTAGES OF EXISTING SYSTEM

- Automatic email filtering may be the most effective method of detecting spam but nowadays spammers can be easily bypass all these spam filtering applications easily.
- Naive Bayes is one of the utmost well-known algorithms applied in these procedures. However, rejecting sends essentially dependent on the context examination can be a difficult issue in the event of bogus positives. Regularly clients and organizations would not need any legitimate messages to be lost.The boycott approach has been probably the soonest technique pursued for the separating of spams .The technique is to acknowledge all the senders other than those from the area/electronic mail ids.

2.5 PROBLEM DEFINITION

Understanding the problem is a crucial first step of solving any machine learning problem. In this project, we will explore and understand the process of classifying emails as spam or not spam. This is called spam classification problem.

The reason to do this is simple: by classifying unsolicited and unwanted emails , we can prevent spam messages from creeping into the user's inbox, thereby improving user experience.

2.6 PROPOSED SYSTEM

- This proposed system, a dataset from "kaggle" website is used as a training dataset. The inserted dataset is first checked for duplicates and null values for better performance of the machine. Then, the dataset is split into 2 sub-datsets; say" train dataset" and "test dataset" in the proportion of 70:30. Then the "train" and "test" dataset is then passed as parameters for text -processing.
- In text-processing, punctuation symbols and words that are in the stop words list are removed and returned as clean words.These clean words then passed for "Feature Transform". In feature transform,the clean words which are returned from the text-processing are then used for 'fit'and 'transform'to create a vocabulary for the machine .The dataset is also passed for "hyperparameter tuning" to find optimal values for the classifier to use according to the dataset.
- After acquiring the values from the "hyperparameter tuning", the machine is fitted using those values with the random state. The state of the trained model and features are saved for future used for testing unseen data. Using classifiers from module sklearn in python, the machines are trained using the values obtained from above [7].

2.7ADVANTAGES OF PROPOSED SYSTEM

Ensemble methods on the other hand proven to be useful as they using multiple classifiers for prediction. Nowadays, lots of email are sent and received and it is difficult as our project is only able to test emails using the limited amount of corpus. Our project, thus spam classification is proficient of classifying mails giving the content of the email and not according to the domain names or any other criteria [7].

- Good Efficiency & Greater Accuracy

CHAPTER-3

ANALYSIS

3.1 INTRODUCTION

In this project, a Spam Mail Detection system is proposed will classify the given email as classification algorithm classifies the given email based on the content. Feature extraction and selection plays avital role in the classification. In spam mail detection, email data is collected through the dataset. To obtain the accurate results, data needs to be pre-processed by removing stop words and word tokenization. Logistic Regression algorithm is used to detect the given email is spam or ham.

3.2 HARDWARE AND SOFTWARE DESCRIPTION

3.2.1 HARDWARE REQUIREMENTS

- System: Pentium i3 Processor
- Hard disk: 256 GB
- Monitor: 14'' LED
- Input devices: Keyboard, Mouse
- RAM: 4 GB

3.2.2 SOFTWARE REQUIREMENTS

- Operating System : Windows 10
- Coding Language : Python
- IDE : Jupyter Notebook
- Tool : Streamlit to build a localhost
- Designing UML Diagrams: StarUML

3.3 BLOCK DIAGRAM

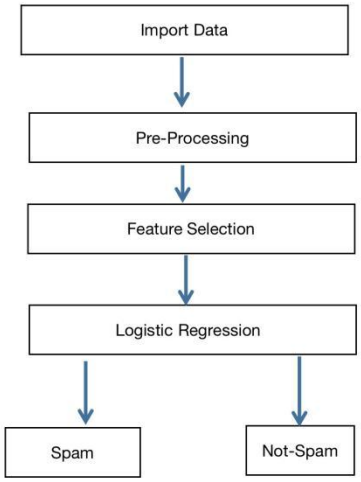

Fig 3.3 Block Diagram of Email spam Classification

3.4 WORKING OF THE PROJECT

3.4.1 DATASET

For this project, the raw data collected is obtained from the online resources kaggle, which is used to train the machine learning models. The data was originally available in English language, and it was obtained in comma separated values (CSV) format. The email data set contain around 5000 emails. Since the approach considered is machine learning approach, the data set is divided into sets for each classifier algorithm. Two sets of email have been prepared. Out of which 80% of the emails will be used to train the models (approximately 4000 emails), whereas remaining 20% will be used to test models individually (1000 emails).

3.4.2 DATA PREPROCESSING

In the machine learning (ML), the preprocessing phrase refers to organizing and managing of raw data before using it to train and test different learning models. In simplistic words, preprocessing is a ML data mining approach that turns raw data into a usable and resourceful structure. The first step in the construction of ML model is preprocessing, in which data from the actual world, typically incomplete, imprecise and inaccurate owing to flaws and deficient is morphed into a precise, accurate,and usable input variables and trends.

3.4.3 LABEL ENCODING

In machine learning, we usually deal with the datasets that contain multiple labels in one or more columns. These labels can be in the form of words or numbers. To make the data understandable or in human-readable form, the training data is often labelled in words.

Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine- readable form . Machine learning algorithm then decide in a better way how those labels must be operated. It is an important pre-processing step for the strutchured dataset in supervised learning. Ino our project we are going to label spam email as 0 and not spam email as 1.

3.4.4 FEATURE EXTRACTION AND SELECTION

Feature extraction is the process of converting the large raw data set into a more manageable format. Any Variable or attribute,or class can be extracted from the data set during the step, depending upon the original data set.

Feature extraction is the crucial step in training the model, which helps in producing more reliable and accurate result.

3.4.5 TRAINING AND EVALUATING THE MODEL

Now that we have to train test split,we would need to choose the model. There is a huge collection of models but for their particular exercise we will be using logistic regression.

Generally when someone asks, what is logistic regression? What do you tell them-Oh! It is an algorithm which is used for categorizing things into two classes (most of time) i.e. the result is mesasured using a dichotonomous variable.But, how does logistic regression classify things like-binomial (2 possible values),multinomial(3 or more possible values) and ordinal(deals with ordered categories).For this post we will only be focusingon binomial logistic regression i.e. the outcome of the model will be categorized into two classes.

CHAPTER-4
DESIGN

4.1 INTRODUCTION

Spam Mails is a well-known practice for sending unwanted or large data to a set of uniqueor random e-mail accounts. Spam Mail A subset of online spam related to the same or identical messages all sent to recipients by email. Spam includes some malware in scripts or other files thatare executable and may harm the user's system. Most e-mail and spam lists are created by scanningthe Usenet ad thoroughly by stealing the Internet email list.

4.2 UML DIAGRAMS

4.2.1 Class Diagram

The class diagram depicts a static view of an application. It represents the types of objects residing in the system ans relationship between them

Contents: Spam Detection, Mapper, Reducer

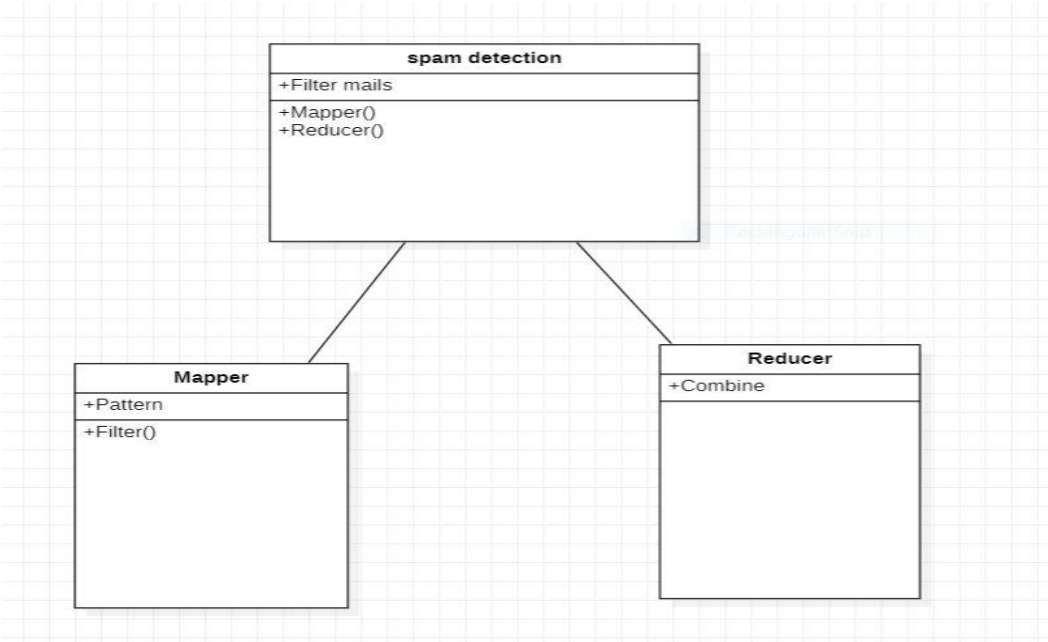

Fig 4.2.1 Class Diagram

4.2.2 Usecase Diagram

A usecase diagram describes a set of interactions between a actor and the system in order to achieve a particular goal

Contents: User , System

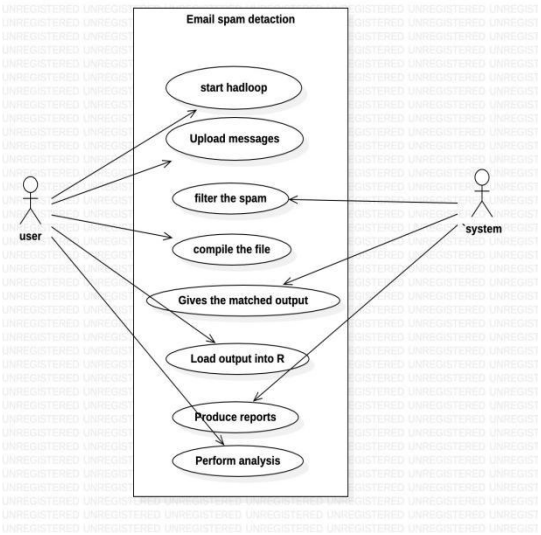

Fig 4.2.2 Use case Diagram

4.2.3 State Chart Diagram

The state chart diagram itself explains the reason for the diagram and different subtitles. It depicts the various states of a part in a system. The states are explicit to a part/object of a system. A state chart diagram depicts a state machine. A state machine is a machine that characterizes various states of a substance and controls these states through outer or interior occasions as displayed in figure 4.2.3

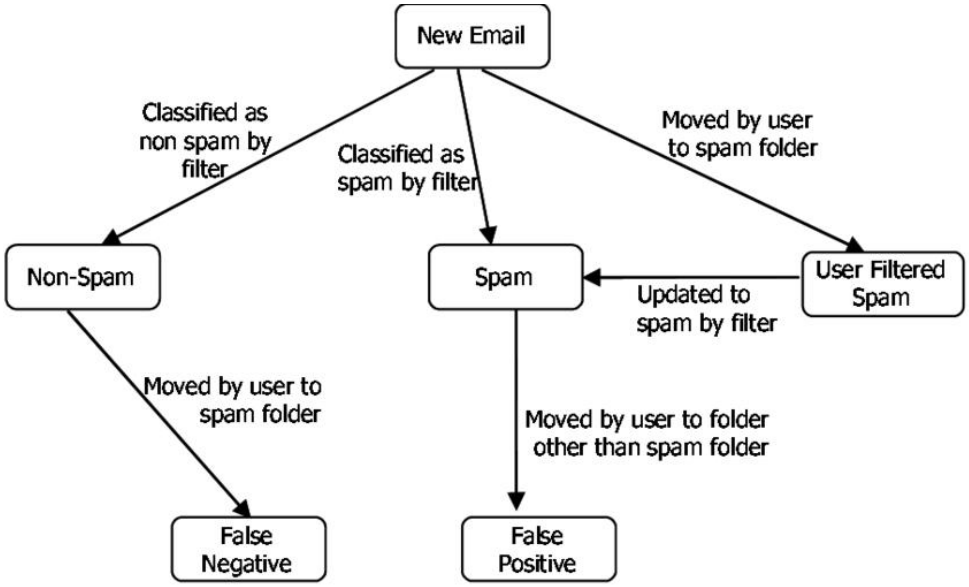

Fig 4.2.3 State Chart Diagram

4.2.4 ACTIVITY DIAGRAM

An activity diagram is a flowchart of activities, as it represents the workflow among various activities. They are identical to the flowcharts, but they themself are not exactly the flowchart.

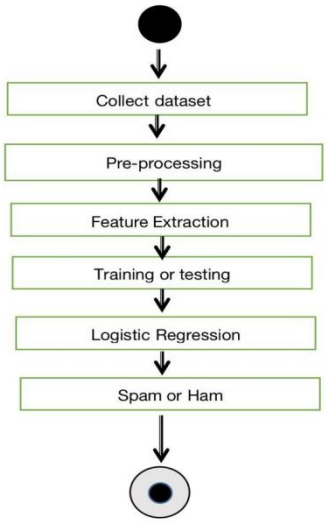

Fig 4.2.4 Activity Diagram

4.3 MODULE DESIGN AND ORGANIZATION

Model design is brief overall description of this framework. The first step is to import all the packages that are used in this project. Then we will remove all stop words and punctuation and duplicate email in preprocessing . In next step we will divide data into training and testing models. Then we will apply logistic Regression algorithm and train data.

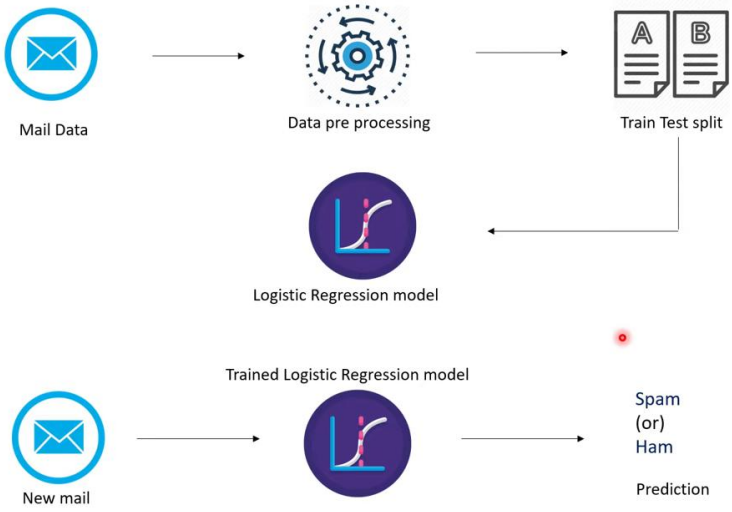

Fig 4.3 Module design

CHAPTER-5
IMPLEMANTATION AND RESULTS

5.1 INTRODUCTION TO LOGISTIC REGRESSION

Logistic regression is one of the most likely and appropriate algorithm used for classification of datasets. In case of classifying a dataset named as spam base the logistic regression is the most versatile decision based approach for detecting spam mails in the dataset. Logistic regression performs some basic test on the given distribution of data which involves finding and calculating some statistical domains like mean and standard deviation. It also produces results of operations like word and character count,max and min operations.After producing and provisioning the statistical and count tests the logistic regression algorithm fetches the outcome of the tests and tends to inter-relate the outcomes.

5.2 WORKING OF LOGISTIC REGRESSION

'Sigmoid function' or 'Logistic function' is implemented as a cost function in Logistic Regression.Hence , for predicting values of probabilistic, the sigmoid functions can be used.

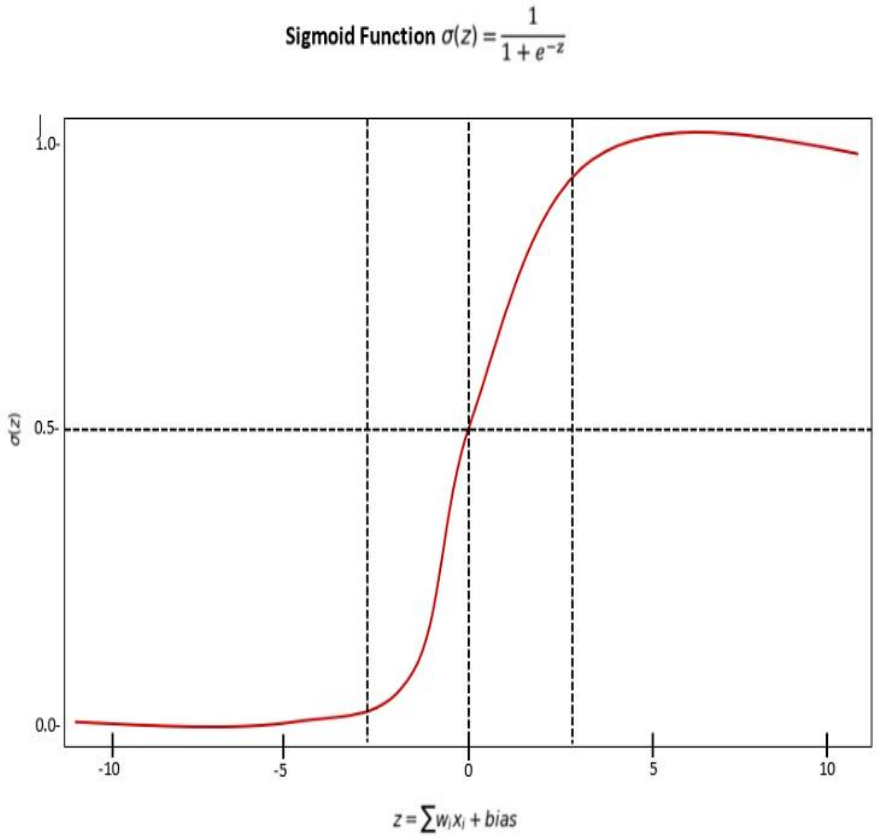

Fig 5.2 Sigmoid Function

First of all, let's have a look at the mathematical equation of the sigmoid function which has been provided below.

$$F(z) = \frac{1}{1-e^{-z}}$$

Now, in the given equation,

$$z = w_0 + w_1.x_1 + w_2.x_2 + \cdots + w_n.x_n$$

As presented in the above equation ,w0,w1,w2,…,wn, is used to represent the regression of the co-efficient of the model that is obtained through maximum Likelihood estimation and xo,x1 x2,……,xn, is used to represent the features or the independent variables. Finally, in the above equation,F(z) calculates the binary outcome probability where the probabilities are classifies according to the provided data point(x) into the two categories.

Advantages of logistic regression :

- Logistic regression is easier to implement, interpret, and very efficient to train.
- It makes no assumptions about distributions of classes in feature space.
- It can easily extend to multiple classes and a natural probabilistic view of class Predictions.
- It is very fast at classifying unknown records.
- Good accuracy for many simple data sets and it performs when the data set is linearly separable.

Disadvantages of logistic regression :

- If the number of observations is lesser than a number of features, Logistic Regression should not be used, otherwise, it may lead to over fitting.
- It constructs linear boundaries.
- Logistic Regression requires average or no multi col-linearity between independent values.
- It can be used to predict discrete functions. Hence, the dependent variable of Logistic Regressions bound to the discrete number set.

5.3 INTRODUCTION TO STREAMLIT

Streamlit is a free and open-source framework to rapidly build and share beautiful machine learning and data science web apps. It is a python library specifically designed for machine learning engineers . Data scientists or machine learning engineers are not web developers and they are not interested in spending weeks learning to use the frameworks to build web apps. Instead, they want a tool that is easier to learn and to use as long as it can display data and collect needed parameters for modeling.

Streamlit allows you to create a stunning- looking application with only few lines of code.

5.4 METHOD OF IMPLEMANTATION

Model Code:

5.4.1 IMPORTING THE DEPENDENCIES

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

5.4.2 DATA COLLECTION & PREPROCESSING

```
# loading the data from csv file to a pandas Dataframe
raw_mail_data = pd.read_csv('mail.csv')
```

```
print(raw_mail_data)
```

```
     Category                                            Message
0         ham  Go until jurong point, crazy.. Available only ...
1         ham                      Ok lar... Joking wif u oni...
2        spam  Free entry in 2 a wkly comp to win FA Cup fina...
3         ham  U dun say so early hor... U c already then say...
4         ham  Nah I don't think he goes to usf, he lives aro...
...       ...                                                ...
5567     spam  This is the 2nd time we have tried 2 contact u...
5568      ham              Will ü b going to esplanade fr home?
5569      ham  Pity, * was in mood for that. So...any other s...
5570      ham  The guy did some bitching but I acted like i'd...
5571      ham                        Rofl. Its true to its name

[5572 rows x 2 columns]
```

```
# replace the null values with a null string
mail_data = raw_mail_data.where((pd.notnull(raw_mail_data)),'')
```

```
# printing the first 5 rows of the dataframe
mail_data.head()
```

	Category	Message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...

```
# checking the number of rows and columns in the dataframe
mail_data.shape
```

```
(5572, 2)
```

5.4.3 LABEL ENCODING

```
# label spam mail as 0;  ham mail as 1;

mail_data.loc[mail_data['Category'] == 'spam', 'Category',] = 0
mail_data.loc[mail_data['Category'] == 'ham', 'Category',] = 1
```

spam - 0

ham - 1

```
# separating the data as texts and label

X = mail_data['Message']

Y = mail_data['Category']
```

```
print(X)
```

```
0       Go until jurong point, crazy.. Available only ...
1                           Ok lar... Joking wif u oni...
2       Free entry in 2 a wkly comp to win FA Cup fina...
3       U dun say so early hor... U c already then say...
4       Nah I don't think he goes to usf, he lives aro...
                              ...
5567    This is the 2nd time we have tried 2 contact u...
5568                 Will ü b going to esplanade fr home?
5569    Pity, * was in mood for that. So...any other s...
5570    The guy did some bitching but I acted like i'd...
5571                           Rofl. Its true to its name
Name: Message, Length: 5572, dtype: object
```

```
print(Y)
```

```
0       1
1       1
2       0
3       1
4       1
       ..
5567    0
5568    1
5569    1
5570    1
5571    1
Name: Category, Length: 5572, dtype: object
```

5.4.4 SPLITTING THE DATA INTO TRAINING & TESTING DATA

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=3)
```

```
print(X.shape)
print(X_train.shape)
print(X_test.shape)
```

```
(5572,)
(4457,)
(1115,)
```

5.4.5 FEATURE EXTRACTION

```
# transform the text data to feature vectors that can be used as input to the Logistic regression

feature_extraction = TfidfVectorizer(min_df = 1, stop_words='english', lowercase='True')

X_train_features = feature_extraction.fit_transform(X_train)
X_test_features = feature_extraction.transform(X_test)

# convert Y_train and Y_test values as integers

Y_train = Y_train.astype('int')
Y_test = Y_test.astype('int')
```

```
print(X_train)
```

```
3075                 Don know. I did't msg him recently.
1787    Do you know why god created gap between your f...
1614                          Thnx dude. u guys out 2nite?
4304                                       Yup i'm free...
3266    44 7732584351, Do you want a New Nokia 3510i c...
                              ...
789     5 Free Top Polyphonic Tones call 087018728737,...
968     What do u want when i come back?.a beautiful n...
1667    Guess who spent all last night phasing in and ...
3321    Eh sorry leh... I din c ur msg. Not sad alread...
1688    Free Top ringtone -sub to weekly ringtone-get ...
Name: Message, Length: 4457, dtype: object
```

```
print(X_train_features)
```

```
  (0, 5413)     0.6198254967574347
  (0, 4456)     0.4168658090846482
  (0, 2224)     0.413103377943378
  (0, 3811)     0.34780165336891333
  (0, 2329)     0.38783870336935383
  (1, 4080)     0.18880584110891163
  (1, 3185)     0.29694482957694585
  (1, 3325)     0.31610586766078863
  (1, 2957)     0.3398297002864083
  (1, 2746)     0.3398297002864083
  (1, 918)      0.22871581159877646
  (1, 1839)     0.2784903590561455
  (1, 2758)     0.3226407885943799
  (1, 2956)     0.33036995955537024
  (1, 1991)     0.33036995955537024
  (1, 3046)     0.2503712792613518
  (1, 3811)     0.17419952275504033
  (2, 407)      0.509272536051008
  (2, 3156)     0.4107239318312698
  (2, 2404)     0.45287711070606745
  (2, 6601)     0.6056811524587518
  (3, 2870)     0.5864269879324768
  (3, 7414)     0.8100020912469564
  (4, 50)       0.23633754072626942
  (4, 5497)     0.15743785051118356
  :     :
(4454, 4602)  0.2669765732445391
(4454, 3142)  0.32014451677763156
(4455, 2247)  0.37052851863170466
(4455, 2469)  0.35441545511837946
(4455, 5646)  0.33545678464631296
(4455, 6810)  0.29731757715898277
(4455, 6091)  0.23103841516927642
(4455, 7113)  0.30536590342067704
(4455, 3872)  0.3108911491788658
(4455, 4715)  0.307141447588111196
(4455, 6916)  0.19636985317119715
(4455, 3922)  0.31287563163368587
(4455, 4456)  0.24920025316220423
(4456, 141)   0.292943737785358
(4456, 647)   0.30133182431707617
(4456, 6311)  0.30133182431707617
(4456, 5569)  0.4619395404299172
(4456, 6028)  0.21034888000987115
(4456, 7154)  0.24083218452280053
(4456, 7150)  0.3677554681447669
(4456, 6249)  0.17573831794959716
(4456, 6307)  0.2752760476857975
(4456, 334)   0.2220077711654938
(4456, 5778)  0.16243064490100795
(4456, 2870)  0.31523196273113385
```

5.4.6 TRAINING THE MODEL

Logistic Regression

```
model = LogisticRegression()
```

```
# training the Logistic Regression model with the training data
model.fit(X_train_features, Y_train)
```

```
LogisticRegression()
```

5.4.7 EVULATING THE TRAINED MODEL

```
# prediction on training data

prediction_on_training_data = model.predict(X_train_features)
accuracy_on_training_data = accuracy_score(Y_train, prediction_on_training_data)
```

```
print('Accuracy on training data : ', accuracy_on_training_data)
```

```
Accuracy on training data :  0.9670181736594121
```

```
# prediction on test data

prediction_on_test_data = model.predict(X_test_features)
accuracy_on_test_data = accuracy_score(Y_test, prediction_on_test_data)
```

```
print('Accuracy on test data : ', accuracy_on_test_data)
```

```
Accuracy on test data :  0.9659192825112107
```

5.4.8 BUILDING THE PREDICTIVE SYSTEM

```
input_mail = ["I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise.

# convert text to feature vectors
input_data_features = feature_extraction.transform(input_mail)

# making prediction

prediction = model.predict(input_data_features)
print(prediction)


if (prediction[0]==1):
  print('Ham mail')

else:
  print('Spam mail')
```

```
[1]
Ham mail
```

5.4.9 CODE FOR BUILDING STREAMLIT WEB APP

`

```
import streamlit as st
import pickle
from sklearn.feature_extraction.text import CountVectorizer
import numpy as np
from win32com.client import Dispatch

def speak(text):
    speak=Dispatch(("SAPI.SpVoice"))
    speak.Speak(text)


model = pickle.load(open('spam.pkl','rb'))
cv=pickle.load(open('vectorizer.pkl','rb'))


def main():
    st.title("Email Spam Classification Application")
    st.write("Build with Streamlit & Python")
    activites=["Classification","About"]
    choices=st.sidebar.selectbox("Select Activities",activites)
    if choices=="Classification":
        st.subheader("Classification")
        msg=st.text_input("Enter a text")
        if st.button("Process"):
            print(msg)
            print(type(msg))
            data=[msg]
            print(data)
            vec=cv.transform(data).toarray()
            result=model.predict(vec)
            if result[0]==0:
                st.success("This is Not A Spam Email")
                speak("This is Not A Spam Email")
            else:
                st.error("This is A Spam Email")
                speak("This is A Spam Email")
main()
```

5.5 OUTPUT SCREEN

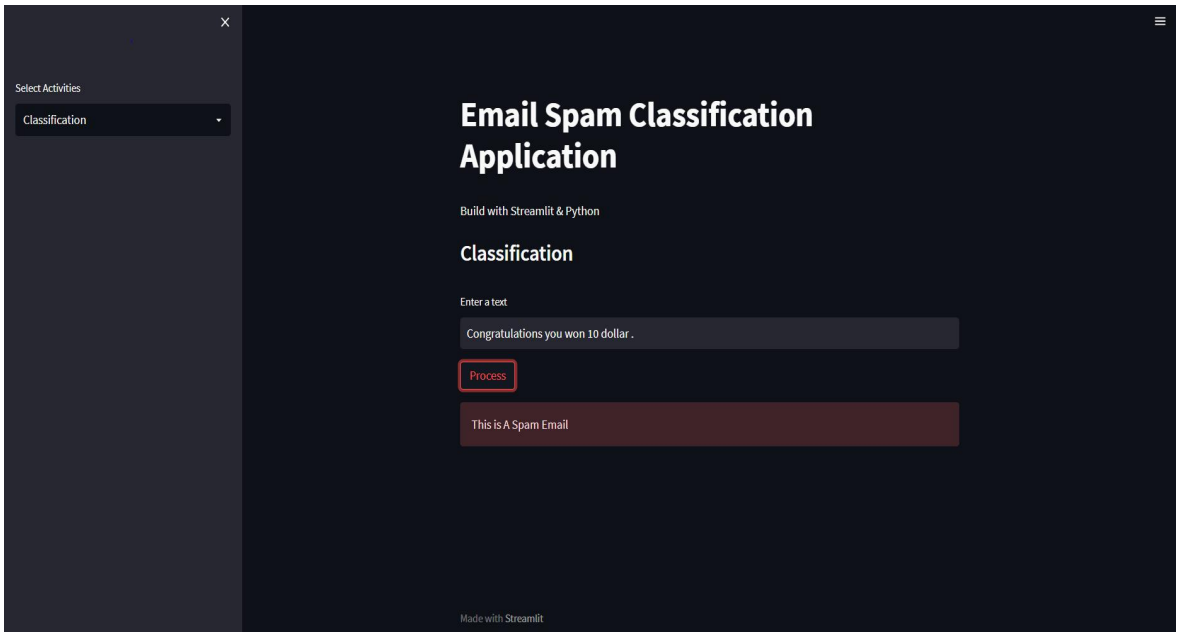

Fig5.5 Output screen of email spam classifier

5.6 RESULT AND ANALYSIS

In Email Spam Classification We use logistic regression because Logistic regression estimates probability, the output will be a number 0 to1; the dependent variable is in binary form. In the case of linear regression, the dependent variable (response variable is continious). We applied logistic regression model for Email Spam Classification model and we got accuracy about 96%

CHAPTER-6
TESTING AND VALIDATION

6.1 INTRODUCTION TO TESTING

Testing is a process, which uncovers the mistakes in the program. It is the significant quality measure utilized during the product advancement measure. During testing, the program executed with a bunch of experiments and the yield of the program for the experiments is assessed to decide whether the program is proceeding as it is required to perform or not.

To ensure that the system doesn't have mistakes, the various degrees of testing technique are applied at contrasting periods of programming advancement.

6.2 DESIGN OF TEST CASE AND SCENARIOS

6.2.1 USER INTERFACE OF EMAIL SPAM CLASSIFICATION

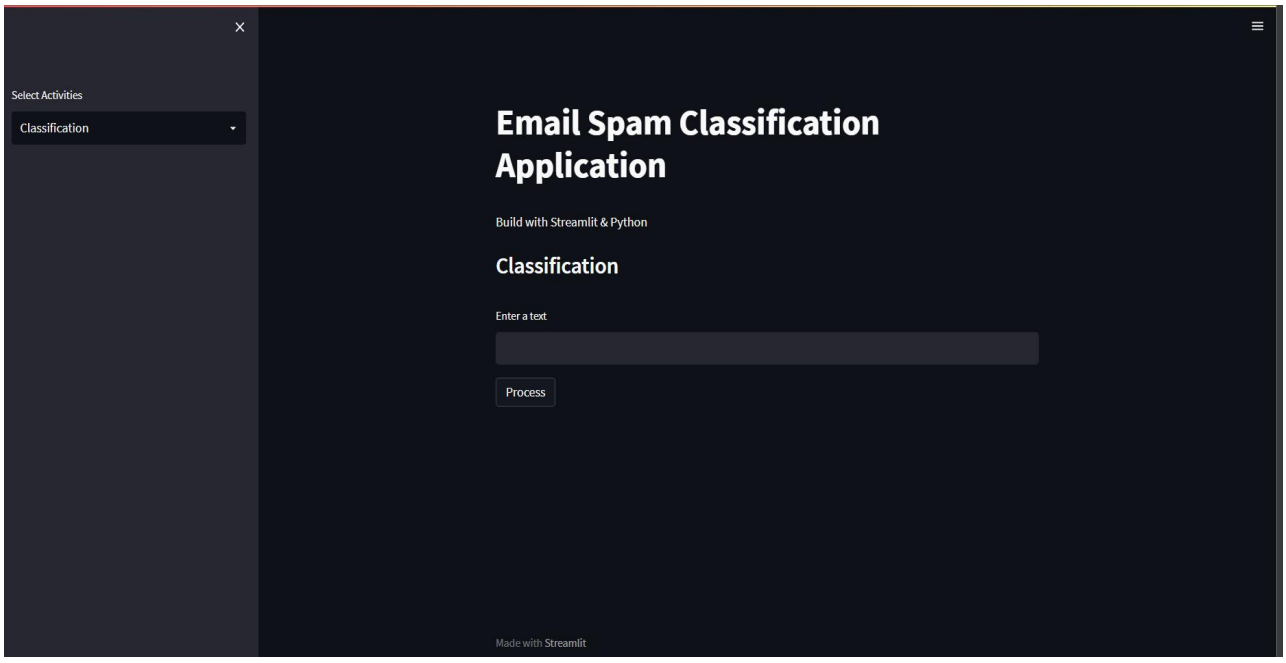

Fig 6.2.1 User Intercace of Email spam classification Application

6.2.2 TESTING FOR SPAM EMAILS

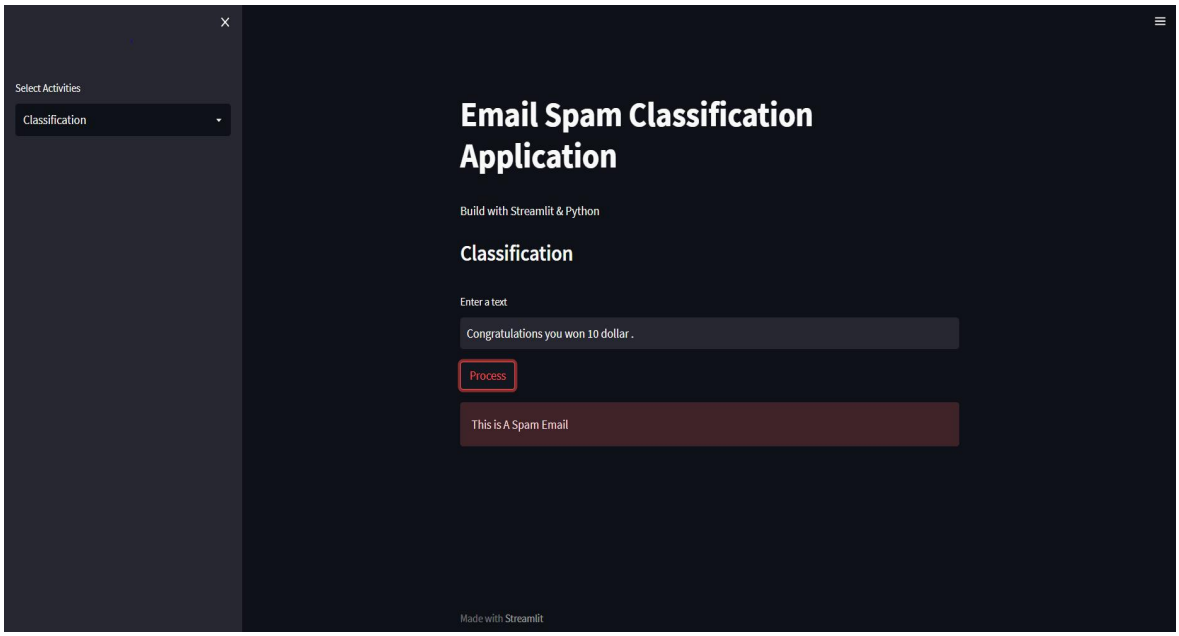

Fig 6.2.2 Testing for spam emails

6.2.3 TESTING FOR NON-SPAM EMAILS

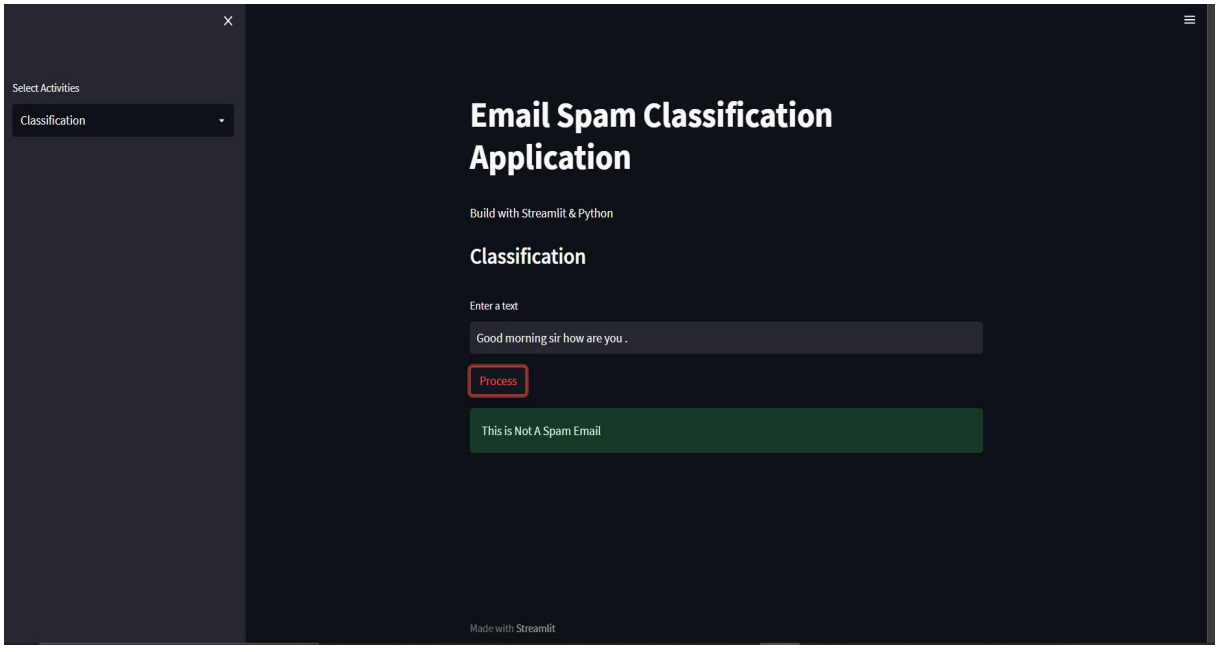

Fig 6.2.3 Testing for non-spam Emails

6.4 CONCLUSION

Spam email is a kind of commercial advertising which is economically viable because email could be very cost effective medium of sender with this . With this Proposed model the specified message can be stated as spam or not using Logistic Regression . We are also building a machine learning web app using streamlit for checking either the message format is spam or not spam.This will help to check the every type of message format.

CHAPTER-7
CONCLUSION

7.1 CONCLUSION

In terms of the number of spam emails sent daily and the number of money people loose everyday because of these spam scams, Spam classification becomes the primary need for all email-providing companies. This article discussed the complete process of spam email classification using advanced technologies of machine learning. We also have closed one possible way of implementing our own spam classifier using one of the most famous algorithm, Logistic regression . We also discussed the case studies of famous companies like Gmail ,Outlook, and Yahoo to review how they use ML and AI techniques to classify such spammers.

7.2 FUTURE WORK

We are expecting that our attempt to understand and implement an accurate method to eradicate spam email will produces best quality results and will help the others to consider our project to their models for producing new products that help the society's protection

We are planning to make an app or software that helps for email spam detection and hoping to make more changes in it.

REFERENCES

REFERENCES

[1] N. Kumar, S. Sonowal, and Nishant, “Email spam detection using machine learning algorithms,” in Proceedings of the 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 108–113, IEEE, Coimbatore, India, July 2020.

[2] A. Zamir, H. U. Khan, W. Mehmood, T. Iqbal, and A. U. Akram, “A feature-centric spam email detection model using diverse supervised machine learning algorithms,” The Electronic Library, vol. 38, no. 3, 2020.

[3] G. Jain, M. Sharma, and B. Agarwal, “Optimizing semantic lstm for spam detection,” International Journal of Information Technology, vol. 11, no. 2, pp. 239–250, 2019.

[4] F. Masood, G. Ammad, A. Almogren et al., “Spammer detection and fake user identification on social networks,” IEEE Access, vol. 7, pp. 68140–68152, 2019.

[5] S. Suryawanshi, A. Goswami, and P. Patil, “Email spam detection: an empirical comparative study of different ml and ensemble classifiers,” in Proceedings of the 2019 IEEE 9th International Conference on Advanced Computing (IACC), pp. 69–74, IEEE, Tiruchirappalli, India, Dec 2019.

[6] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab, “A comprehensive survey for intelligent spam email detection,” IEEE Access, vol. 7, pp. 168261–168295, 2019

[7] Nikhil kumar,Sanket Sonowal, Nishant “ Email Spam Detection using Machine Learning Algorithms”. IEE CONFERENCE 2020.