Spam Mail Prediction System using Machine Learning

A project report
Submitted in partial fulfilment of the requirements for the award of the degree of

Master of Computer Applications
In
Computer Science

By

PRINS VERMA

MCA VI Semester

Enrolment No.: U1949059

Under the Supervisions of

Prof. T.J. Siddiqui

Department of Electronics and Communication
J K Institute of Applied Physics & Technology
UNIVERSITY OF ALLAHABAD
Prayagraj – 211002, India

May 20, 2022

CANDIDATE'S DECLARATION

I, **Prins Verma,** hereby certify that the work, which is being presented in the report/thesis, entitled "**Spam Mail Prediction System Using Machine Learning"**, in partial fulfillment of the requirement for the award of the Degree of **Masters of Computer Application** and submitted to the institution is an authentic record of my own work carried out during the period Feb, 2022 to May, 2022 under the supervision of **Prof. T.J. Siddiqui** at the Department of Electronics and Communication, University of Allahabad. The matter presented in this report/thesis has not been submitted elsewhere for the award of any other degree or diploma from any Institutions.

I declare that I have cited the reference about the text(s) /figure(s) /table(s) /equation(s) from where they have been taken. I further declare that I have not willfully lifted up some other's work, para, text, data, results, etc. reported in the journals, books, magazines, reports, dissertations, theses, etc., or available at web-sites and included them in this report/thesis and cited as my own work.

Date: May 20, 2022 Signature of the Candidate

Prins Verma

CERTIFICATE FROM THE SUPERVISOR

This is to certify that the **Mr. Prins Verma** has carried out this project/dissertation entitled **Spam Mail Prediction System Using Machine Learning** under my supervision.

Date:

Signature of the Supervisor
Prof. T.J. Siddiqui
(Supervisor)
Seal/Designation

ACKNOWLEDGEMENT

Project is sort of a bridge between theoretical and practical working. With this willing I joined this particular project. I am feeling oblige in taking the chance to sincerely thanks to **Prof. T.J. Siddiqui Ma'am** for his guidance, generous attitude, inspiration and constructive suggestions that help me in the preparation of this project. I am also thankful to all my friends who have always helping and encouraging me though out in successful completion of the project.

Prins Verma

ABSTRACT

This project is based on study conducted on predicting spam mail from emails. Email is the most popular way to interact with others through the internet nowadays. Email is used to communicate with both ways formal and informal.

As we know, use of emails is increasing rapidly. Along with these mails, Spam mails are also sent through different platforms and these spam mails are difficult to identify that it is a spam or ham mail. Spam mail consumes almost 90% of billions of emails sent every day.

Spam mails are junk mails which are sent by a spammers or hackers through mail to target a user's identity or stole his personal data. As a result, it's critical to determine whether spam emails are fraudulent. This project will identify those spam mails using machine learning techniques. This project will discuss machine learning algorithms: Logistic Regression Algorithm and apply it to our data sets. The precision and accuracy of the method must be calculated.

Keywords: Email, Machine Learning, Logistic Regression, Security, Prediction

TABLE OF CONTENTS

Title Page No.

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND AND MOTIVATION

- Although we now have numerous popular chat services such as WhatsApp, Facebook Messenger, and Snapchat, e-mail remains a vital component of our everyday digital lives. According to a research, global e-mail subscribers are expected to increase to 4.6 billion in 2025, up from 4.258 billion in 2022. Emails employ end-to-end encryption to prevent unauthorized users from reading other people's messages, as the message is only visible to the sender and receiver. Companies can exchange detailed information via email attachments such as spreadsheets, word reports, photos, audio, and video. [1]
- Email is ubiquitous, whether in business or in our personal lives. It is currently the most important component of everyone's life. Email is used for both formal (business/office) and casual (personal relationship) communication. It's all over the place. As a result, according to Statista, almost 281.1 billion emails are sent every day around the world. That means nearly 37 emails are sent every person on the earth, and it's alarming to learn that more than half of those emails are spam, lowering productivity and exposing users to phishing and cyber-attacks. [2]
- Spam e-mail is simply unsolicited mail or junk mail sent by untrustworthy individuals or businesses for their own personal gain, such as business promotion or data collection. Spam is often sent for commercial purposes. Botnets, or infected computer/system networks, can send it in massive amounts.
- Spam is increasingly being seen as a more serious message danger, as it is increasingly being used to send worms, viruses, and Trojans, as well as more directly financial rooks.
- As a result, every user should have strong security software in place and use extreme caution while opening e-mails.

1.2 PROBLEM STATEMENT AND OBJECTIVES

- Spam email can cause a number of issues, including:
  - Spam prevents the user from utilizing full and good use of their time,storage capacity and network bandwidth.
  - The huge volume of spam e-mails flows through the computer networks have destructive effects on the memory space of e-mail servers, communication bandwidth, Central Processing Unit (CPU) power and user time.
  - The threat of spam e-mail is on the increase on yearly basis and it is responsible for over 77% of the whole global e-mail traffic.
  - It is also resulted to untold financial loss to many users who have targeted victim of internet scams and other fraudulent practices of spammers who send e-mails pretending to be from reputable companies (like: Apple, Samsung etc) with the intention to persuade individuals to disclose sensitive personal information like passwords, Bank Verification Number and credit card numbers.

- The main objective of this project is to design a system which filters all spam emails entered by a user and to do that I am going to use some machine learning algorithm. Such as: **Logistic Regression Model**
- Based on this trained model, all the emails can be filtered in two parts: Spam and Ham mail. Where a user can easily identify Spam mail and prevent himself from it. The result obtained from this project shows that it has high accuracy (99.69%).

1.3 PROJECT CATEGORY

The project category of **Spam Mail Prediction System** is Machine Learning, fraud detection, filtration, probability and data analysis.

1.4 TOOLS/PLATFORM REQUIREMENT

- A computer system having minimum hardware requirement:-
  1. Processor: Intel Core i3 or above
  2. RAM: 4GB or above
  3. Operating System: Windows 7 or above
  4. Disk Size: 500GB

- Software Requirements
    1. Python Version: 3.9.X
    2. Google Colab
    3. Machine Learning

1.5 CONTRIBUTION OF THE THESIS/REPORT

In this project my aim is to design a improved, updated and effective model with the help of machine learning algorithm that will perform its calculations on the basis of previous data which is being provided and it will predict whether entered mail is Spam or Ham. Based on the prediction given by the model we can easily identify that the mail is Spam or Ham as well as we can prevent fraud or scam.

1.6 STRUCTURE OF THE THESIS/REPORT

Chapter 1, introduced about the spam mail Prediction System, its background and motivation, how spam mail is used to fraud / steal personal data of a person, just by click on a link provided in a mail. Spam mails are also point of concerned because these mails are waste of memory and consume user time. I also made a small contribution in this project by updating all current data by the help of reading various thesis/report.

Chapter 2, described the previous works done by others in this particular area. I considered the works analyze their merits and demerits, and shortcomings that motivate me to work further in this area. This chapter contains a specialized overview of Spam mail Prediction.

Chapter 3, introduced Spam mail Prediction, how its work in easy terms that a normal people can understand, who don’t have any idea of how this Spam Mail Prediction System is working. I have explained all the methodology I have used in this project with a basic work plan by the help of diagrams and algorithm.

Chapter 4, In this chapter I have described step by step implementation of the of my work plan.

Chapter 5, Conclude about Spam Mail Prediction System and its scope for future work.

CHAPTER 2

LITERATURE REVIEW/SURVEY

To comprehend and forecast spam mail, extensive research has been conducted. Many algorithms are being used in the field of email spam filtering, and many studies have generated distinct predictive models based on different machine learning techniques. Deep Learning, Nave Bayes, Support Vector Machines, Neural Networks, K-Nearest Neighbor, Rough Sets, and Random Forests are examples of such methods.

For email filtering, they compared SVM with Decision Tree. They separated the data into two groups: training and testing. Each model is individually taught, and its correctness is assessed based on that training. The author used supervised learning for both methods, and the Support Vector Machine approach had an accuracy of 92 percent and the Decision Tree method had an accuracy of 82 percent. The author determined that the performance of Support Vector Machine is superior than that of Decision Tree based on his research. [3]

The author developed a system for integrating categorization techniques to improve spam mail filtering results. The author used data mining to collect all of the information on spam filtering's success, present challenges, and previous failures. The approach relied on binary categorization, with 1 indicating Spam email and 0 indicating Ham email. For email filtering, they integrated the two methods into one: Machine Learning and Knowledge Engineering. On the combined K-Nearest Neighbor and Support Vector Machine technique, the proposed solution performed poorly. [4]

All of the datasets (spam or ham emails) were gathered by the author from various sources. The author evaluated all of the data and carefully picked 23 criteria that they considered to be deciding factors in whether an email was ham or spam. They assigned a value to each of the criteria and determined a threshold value based on the analysis. The total value of each email for classification was calculated and examined to see if it was greater than or less than the threshold value, and the result was given based on that. Because the study was conducted on a small dataset of approximately 750 emails, this proved ineffective. [5]

For email filtering, a method based on the SVM algorithm and feature extraction was proposed. This approach includes numerous processes, such as Email Collection, in which data from the dataset is gathered. After that, data is sent for pre-processing, which removes any extraneous content and only sends the desired content for processing. The feature extraction technique is then followed by SVM model training. The author used the Apache Public Corpus dataset. Special symbols, HTML tags, URLs, and extraneous alphabets were deleted from the proposed approach. All of the terms from the dictionary were mapped using the Vocab file. An accuracy of 98 percent was reached using the SVM method on a pre-processed dataset.[6]

CHAPTER 3

DESIGN DETAILS, PROPOSED APPROACHES, SYSTEM AND ALGORITHMS

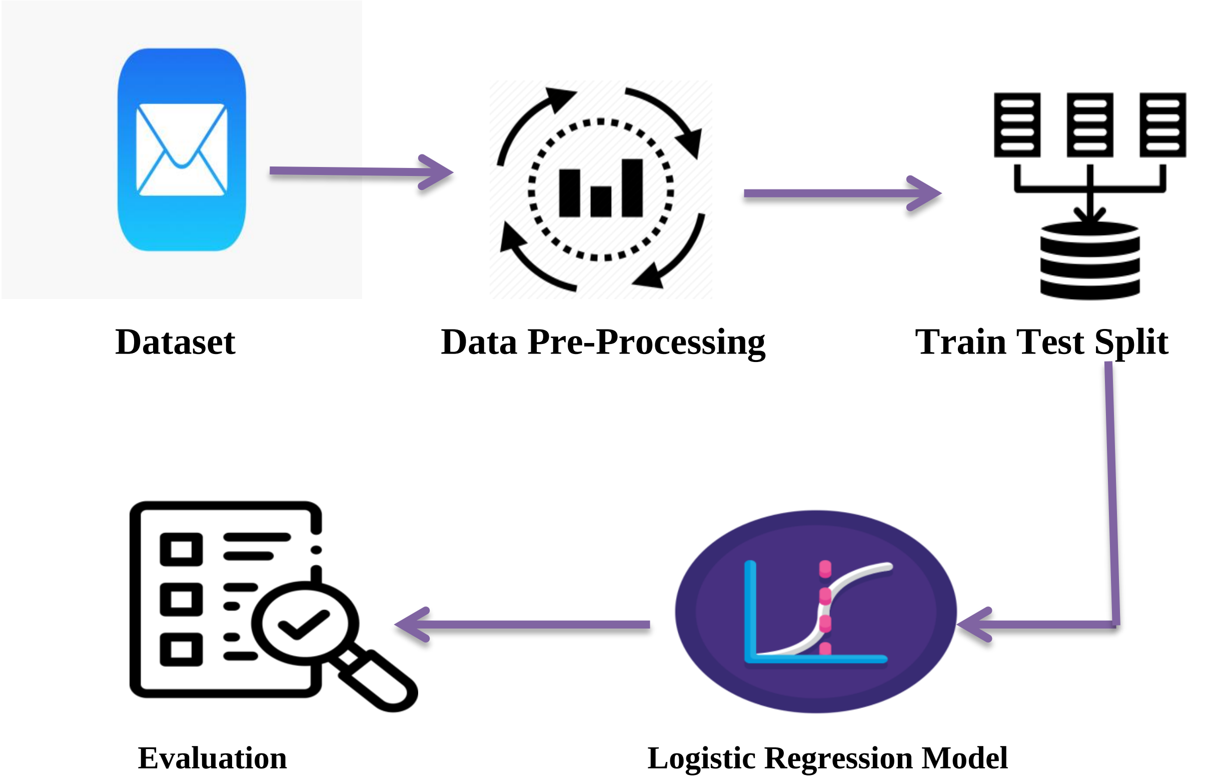
3.1 INTRODUCTION

- This project uses a dataset from the "Kaggle" website as a training dataset. The inserted dataset is initially checked for duplicates and null values to improve machine performance. The dataset is then partitioned into two sub-datasets with an 80:20 ratio, such as "train dataset" and "test dataset." Text-processing settings are then specified for the "train dataset" and "test dataset."
- During text processing, punctuation symbols and terms on the stop words list are removed and replaced with clean words. These clean phrases are then supplied to the "Feature Transform" command. In feature transform, the clean words obtained from text processing are used for 'fit' and 'transform' to create a vocabulary for the computer. The dataset is also utilized for "hyper parameter tuning," which entails using the dataset to determine the best values for the classifier to employ.
- After receiving the values from the "hyper parameter tuning," the machine is fitted using those values with a random state. The state and properties of the trained model are saved for future testing with unknown data. Using classifiers from the Python sklearn, the machines are trained using the values acquired from above.

3.2 REQUIREMENT SPECIFICATION

The purpose of this project is to design a model which predict, whether entered mail is spam or ham. To design this project, we need dataset which is taken from Kaggle website and on the basis of this dataset we are going to train and test our model. On this dataset, logistic Regression Algorithm will be used because this algorithm is best for the binary classification problem. Software used in this project is: Python, Machine Learning (ML) and Google Colab.

3.3 METHODOLOGY USED



- **Dataset:**

The mail objective of the project is to predict the spam mail with high accuracy from others. For this, we have taken raw mail dataset consisting both Spam and Ham mail from "Kaggle.com" website.This raw mail dataset contains **5170 message** having tags as “Spam” and “Ham” where total spam is **1499** and ham is **3672**. This dataset will be used to train machine learning model.

- **Data Pre-processing:**

The dataset had many unnecessary columns so it was removed from the dataset. Now the dataset has only two columns which we need as “Category” and “Message”. We cannot feed this raw data to our machine learning algorithm so we need to pre-process this data properly. In this stage we will remove all the unwanted data like: null values, extra irrelevant information.

In next step, label encoding was done. In label encoding, spam assigned as ‘0’ and spam as ‘1’.

Machine doesn’t understand Natural Language that’s why all the mails were converted in machine readable form i.e. numeric form so that machine can easily understand and our model can have high accuracy.

- **Feature Extraction:**

Because our algorithm expects integers as input, we'll need a feature extraction layer in the middle to convert the words to integers. As we know, it is easier for a machine / computer to understand numbers but it is very tough to understand text and paragraphs, so we will do some processing where we will convert this text into more meaningful numbers and that will be done in this part.

- **Train Test Split:**

The data set will be divided into training and test data. We will use 80% of data to train our machine learning model and 20% data to evaluate our model, so we can find how well our model is performing. We will use test data to test our machine learning model so that we can find its accuracy.

- **Logistic Regression Model:**

In its most basic form, logistic regression is a statistical model that uses a logistic function to model a binary dependent variable. Logistic Regression Model is best when it comes to binary classification problem. There are two classes in it and we are trying to classify these into two classes. The two classes are:

  - Spam mail
  - Ham mail

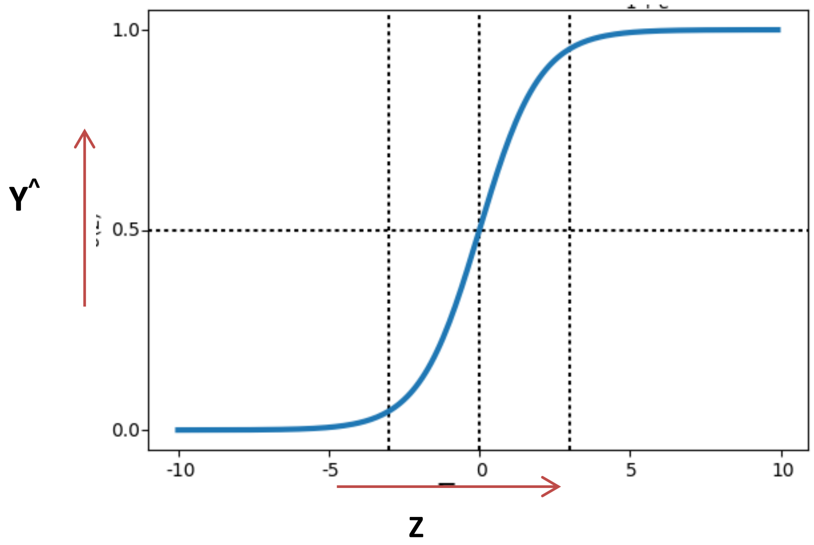
- **Evaluation:**

We will evaluate our model to check where it working properly or not as well as we will also going to check models accuracy.

3.4 ALGORITHM

Logistic Regression: We employ logistic regression as one of the most important machine learning models. A supervised learning model is logistic regression. Supervised learning is something where we use labeled data.

- Logistic regression is a classification model. It is mainly used for classification, where classification is something we just group the data or we classify the data into different classes. As in the project you can see there is two kinds of classes: one set of data is Category and second one is Message and on the basis of these classes we will classify that the upcoming mail is Spam or Ham mail.
- Logistic Regression is best suited for Binary Classification Problem. Binary Classification means when you have two classes or two categories.
- It uses **Sigmoid Function**. Sigmoid is a mathematical function that takes any real number and maps it to a probability between 1 and 0.



$Y^{\wedge}$ - Probability that (Y=1)

$Y^{\wedge}$ = P(Y=1 I X)

X – Input features

w – Weights (Number of weights is equal to the number features in a dataset)

b – bias

$Y^{\wedge} = \sigma(Z)$

$Y^{\wedge} = 1 / (1+e^{-z})$

$Z = w.X + b$

CHAPTER 4

SIMULATOR/TOOL/TECHNOLOGY DETAILS, AND SIMULATION/IMPLEMENTATION

IMPLEMENTATION

In this chapter I am going to discuss step by step implementation of my work plan.

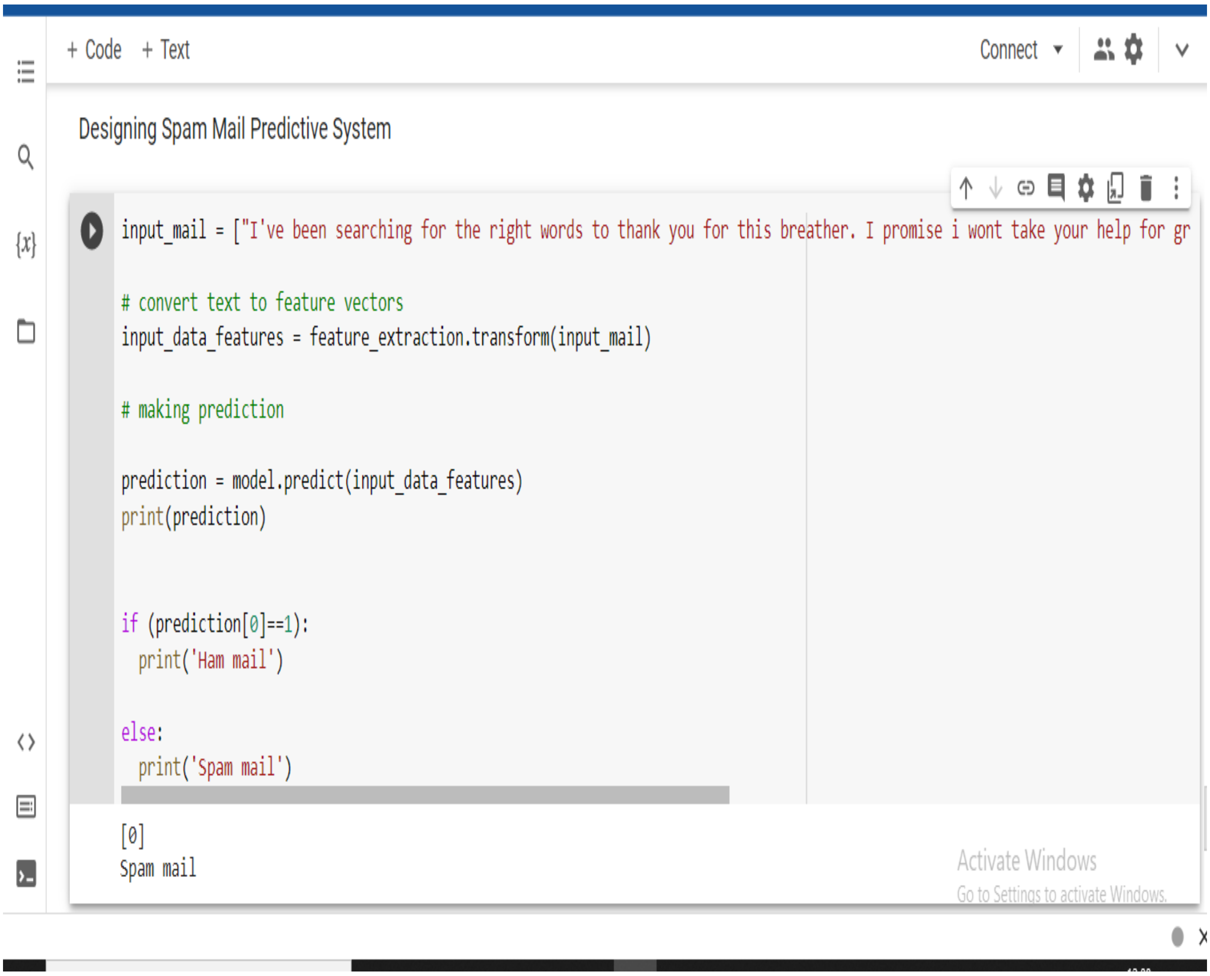
- First, find out the problem and the problem is, "how we can identify that: a mail is spam or ham?"
- We have the problem so think about the solution of this problem and the solution is: we can design a **System** which can predict that, the entered mail is spam or ham.
- We need mail dataset which is available on kaggle website.
- Performing data pre-processing on dataset to eliminate spaces, special character, stickers etc, so that our system works accurately.
- Our algorithm always expects the input to be integers, so we need to have some **feature extraction layer** in the middle to convert the words to integers. As we know, it is easier for a machine / computer to understand numbers but it is very tough to understand text and paragraphs, so we will do some processing where we will convert this text into more meaningful numbers and that will be done in this part.
- Labelling the data as: **'0'** as Spam and **'1'** as Ham
- Splitting our data set into two parts for
  1. Training model and
  2. Testing model.
- The ration of training and testing dataset will be 80:20 i.e. 80% of data will be used for training the model and 20% for testing.
- As there is only two classes (Category and Message), we will use Logistic Regression algorithm because it's best for Binary Classification problems in the term of Accuracy.
- At last evaluate the trained model as well as find out its accuracy.
- Our Spam Mail Prediction System is ready. Give a mail as input and you will see result as "spam mail [0]" or "ham mail [1]".

.

CHAPTER 5

TESTING, VERIFICATION, VALIDATION RESULTS AND DISCUSSIONS

As our model is trained. Its time to discuss its accuracy and result. Total 125 mails were entered during testing phase and we got 124 correct answers as spam and ham mail which is high in comparison with others algorithm as well proposed system.



Designing Spam Mail 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 gr

# 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')
```

```
[0]
Spam mail
```

CHAPTER 6

CONCLUSION AND FUTURE WORK/EXTENSION

6.1 INTRODUCTION

In this chapter will summarize the whole project and conclude the main aim of the project that i.e. training and testing our model as well as I am going to explain the future scope of our model and its real life implementation.

6.2 CONCLUSIONS

In this project, we examined the machine learning approach and its application to the field of spam prediction. Spam emails have become a big source of concern for the internet community because they threaten users' integrity and productivity. Email filtering is critical for effective email communication. The accurate identification of spam emails is a major problem, and several researchers have proposed numerous filtering approaches. The project looked at some of the publicly available datasets and performance indicators that can be used to assess the efficacy of spam prediction systems.

After analyzing different papers given by different researchers, we observed as follows:

- SVM algorithm was not able to give better result in terms of accuracy.
- Decision Tree Classifier was taking large memory space which is a great matter of concern.
- Size of dataset used by many researchers is very small and needs to be expanded.

6.3 SCOPE FOR FUTURE WORK

This project is based on the study of predicting the spam mail to prevent fraud / malicious activities done by a person to steal someone’s personal data or identities.

As we know, spam mails are sent by a spammer. These spammers are aware of spam filter algorithms which are used for filtering spam mails, so they can try another technique to by-pass the filter like: sending a mail containing PDF file, image etc.

To avoid / prevent/ filter these types of mails we can modify our model, like: we can use different dataset which contains PDF file or image.

REFERENCES

[1] Introduction: https://perfectelearning.com/home/details/all/spam- mail-detection-using-machine- learning#:~:text=The%20machine%20learning%20model%20used,evading%20their%20email%20spam%20filter

[2] Introduction: https://www.statista.com/statistics/456500/daily- number-of-e-mails-worldwide/

[3] A.S Yuksel, S.F. Cankaya, I.S Uncu, International Research Journal of Engineering and Technology (IRJET)-2017)

[4] V.K Singh, S. Bhardwaj, International Research Journal of Engineering and Technology (IRJET)-2018)

[5] A.S Aski, N.K Sourati, International Research Journal of Engineering and Technology (IRJET)-2016)

[6] T. Verma, International Research Journal of Engineering and Technology (IRJET) (2017)

Some sample of IEEE style referencing:

[1] D. J. Beebe, “Signal conversion (Book style with paper title and editor),” in *Biomedical Digital Signal Processing*, W. J. Tompkins, Ed. Englewood Cliffs, NJ: Prentice-Hall, 1993, ch. 3, pp. 61–74.

[2] M. Akay, *Time Frequency and Wavelets in Biomedical Signal Processing* (Book style). Piscataway, NJ: IEEE Press, 1998, pp. 123–135.

[3] G. B. Gentili, V. Tesi, M. Linari, and M. Marsili, “A versatile microwave plethysmograph for the monitoring of physiological parameters (Periodical style),” *IEEE Trans. Biomed. Eng.*, vol. 49, no. 10, pp. 1204–1210, Oct. 2002.

[4] V. Medina, R. Valdes, J. Azpiroz, and E. Sacristan, “Title of paper if known,” unpublished.

[5] E. H. Miller, “A note on reflector arrays (Periodical style—Accepted for publication),” *IEEE Trans. Antennas Propagat.*, in press.

[6] T. Menendez, S. Achenbach, W. Moshage, M. Flug, E. Beinder, A. Kollert, A. Bittel, and K. Bachmann, “Prenatal recording of fetal heart action with magnetocardiography” (in German), *Zeitschrift für Kardiologie*, vol. 87, no. 2, pp. 111–8, 1998.

[7] J. E. Monzon, “The cultural approach to telemedicine in Latin American homes (Published Conference Proceedings style),” in *Proc. 3rd Conf. Information Technology Applications in Biomedicine, ITAB´00*, Arlington, VA, pp. 50–53.

[8] F. A. Saunders, “Electrotactile sensory aids for the handicapped (Presented Conference Paper style),” presented at the *4th Annu. Meeting Biomedical Engineering Society*, Los Angeles, CA, 1973.

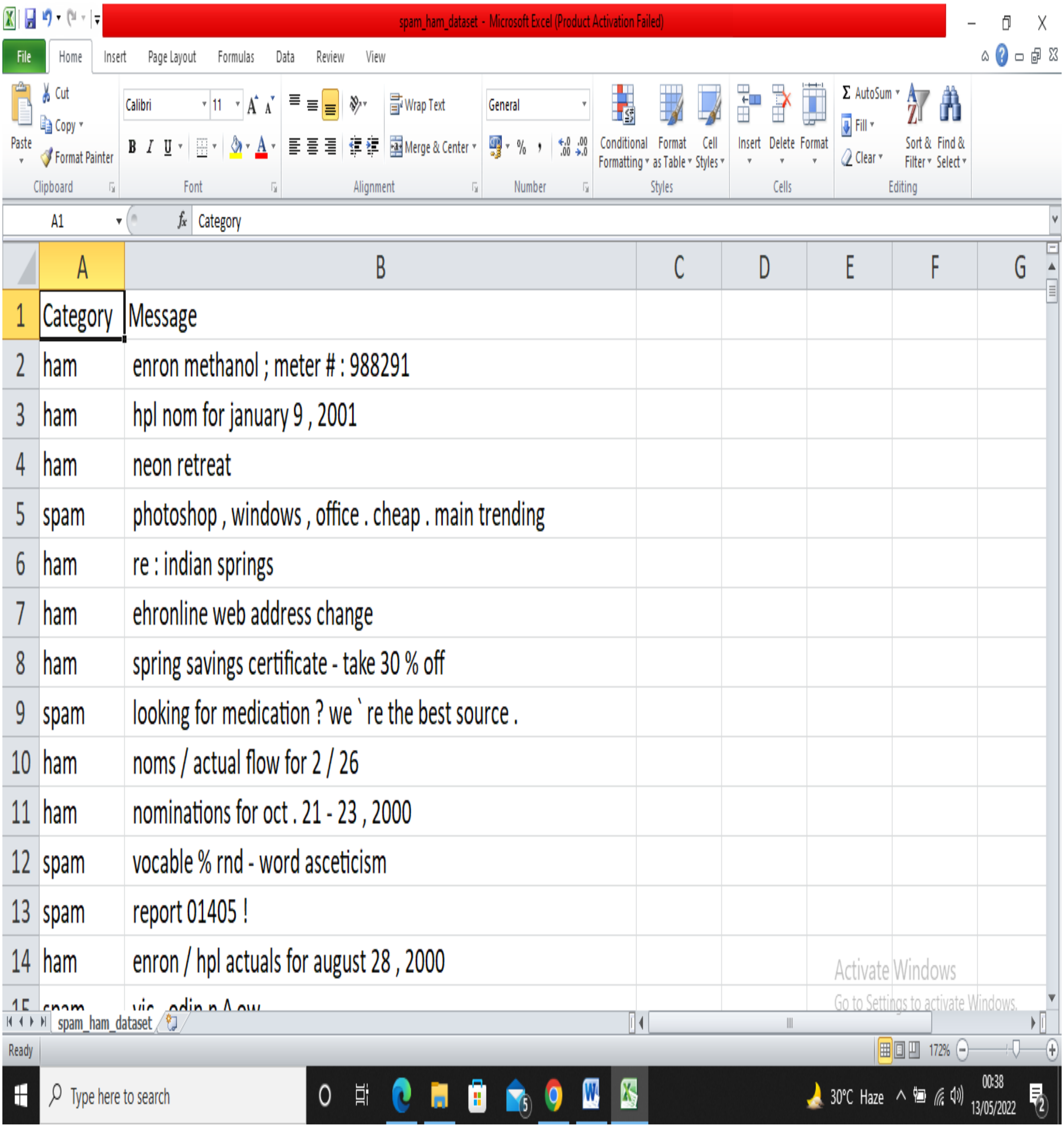
[9] J. R. Boheki, “Adaptive AR model spectral parameters for monitoring neonatal EEG (Thesis or Dissertation style),” Unpublished Ph.D. dissertation, Biomed. Eng. Program, Univ. Fed. Rio de Janeiro, Rio de Janeiro, Brazil, 2000.

[10] Nikhil Kumar, Sanket Sonowal, Nishant “Email Spam Detection Using Machine Learning Algorithms”, IEEE CONFERENCE 2020.

APPENDICES

A.1 ***Dataset:*** Dataset is size large so here is a link to access that dataset:

https://drive.google.com/file/d/1VAHJsrrhHXGC9fq2G3QxMue6gXaVvcNl/view?usp=sharing



	A	B
1	Category	Message
2	ham	enron methanol ; meter # : 988291
3	ham	hpl nom for january 9 , 2001
4	ham	neon retreat
5	spam	photoshop , windows , office . cheap . main trending
6	ham	re : indian springs
7	ham	ehronline web address change
8	ham	spring savings certificate - take 30 % off
9	spam	looking for medication ? we ` re the best source .
10	ham	noms / actual flow for 2 / 26
11	ham	nominations for oct . 21 - 23 , 2000
12	spam	vocable % rnd - word asceticism
13	spam	report 01405 !
14	ham	enron / hpl actuals for august 28 , 2000
15	spam	vic , odin n ^ ow

A.2 *Code:*

Link to access Google Colab file: Code File

Importing Dependencies / Libraries

```
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
```

- **Data Collection and Pre-processing**

```
# loading the data from csv file to a pandas Dataframe
        raw_data = pd.read_csv('/content/drive/MyDrive/Major Project/spam_ham_dataset.csv')
#printing Raw Mail Data
        print(raw_data)
```

- **Replacing all the null values with a null string**

```
        mail_data = raw_data.where((pd.notnull(raw_data)),'')
```

- **Checking Dataset**

```
        # printing the first 5 rows of the dataframe
                mail_data.head()
        # printing the last 5 rows of the dataframe
                mail_data.tail()
        # checking the number of rows and columns in the dataframe
                mail_data.shape
```

- **Assigning Label/ 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
```

Reminder

spam = 0

ham = 1

- **Separating the data**

```
# separating the data as texts and label
        X = mail_data['Message']
        Y = mail_data['Category']
```

- **Verifying after separation**

```
        print(X)
        print(Y)
```

- **Splitting the dataset into two parts:**
    1. **Training data (X_train, Y_train)**
    2. **Testing data (X_test, Y_test)**

```
        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)
```

- **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)
        print(X_train_features)
```

- **Training the model**
    1. **Logistic Regression**

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

- **Evaluation of trained model**

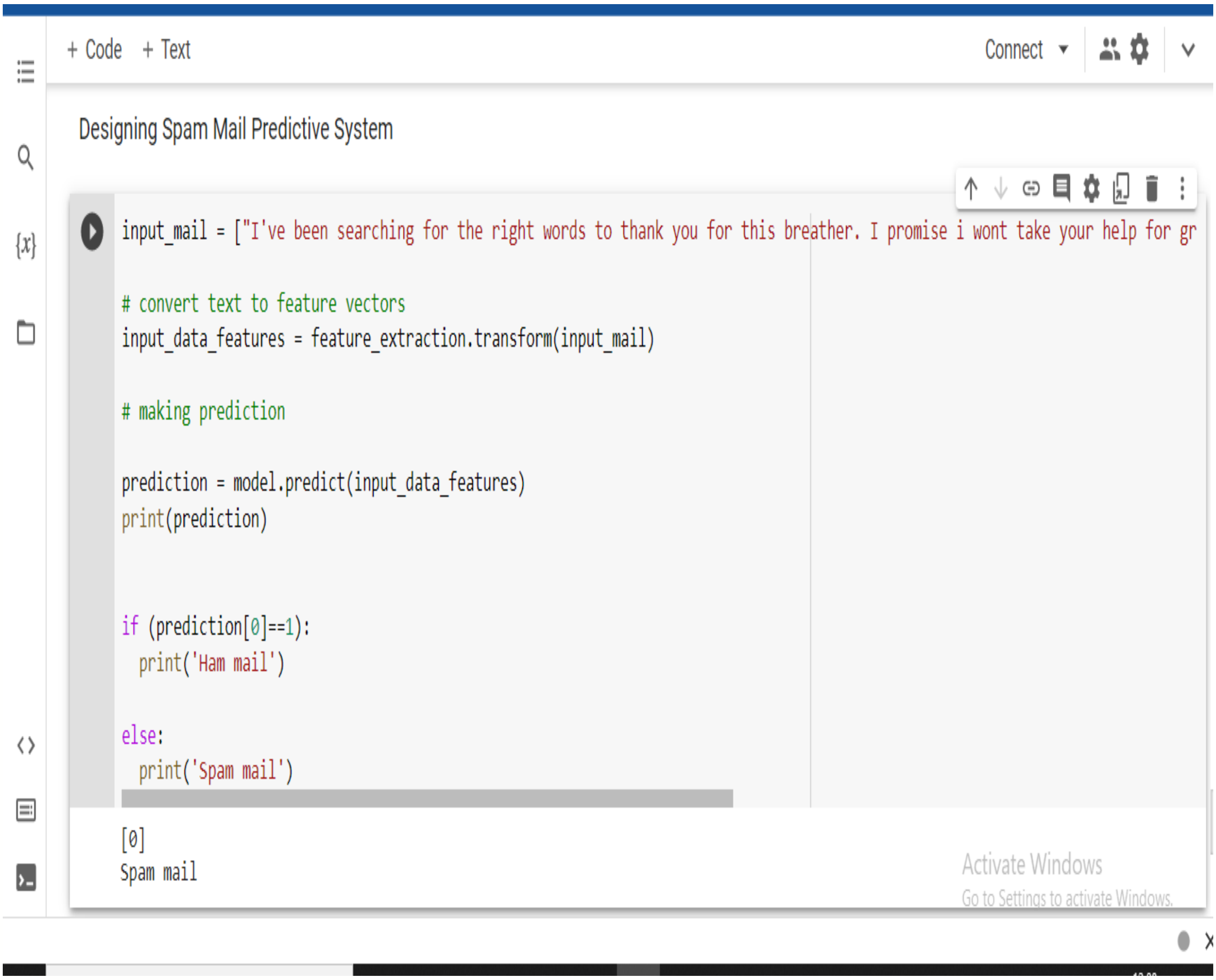
```
# prediction on training data
        prediction_on_training_data = model.predict(X_train_features)
        accuracy_on_training_data = accuracy_score(Y_train, prediction_on_training_data)
        print('Accuracy on training data : ', accuracy_on_training_data)
```

- **Designing Spam Mail Prediction 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. You have been wonderful and a blessing at all times"]

# 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')
```

Screenshot:



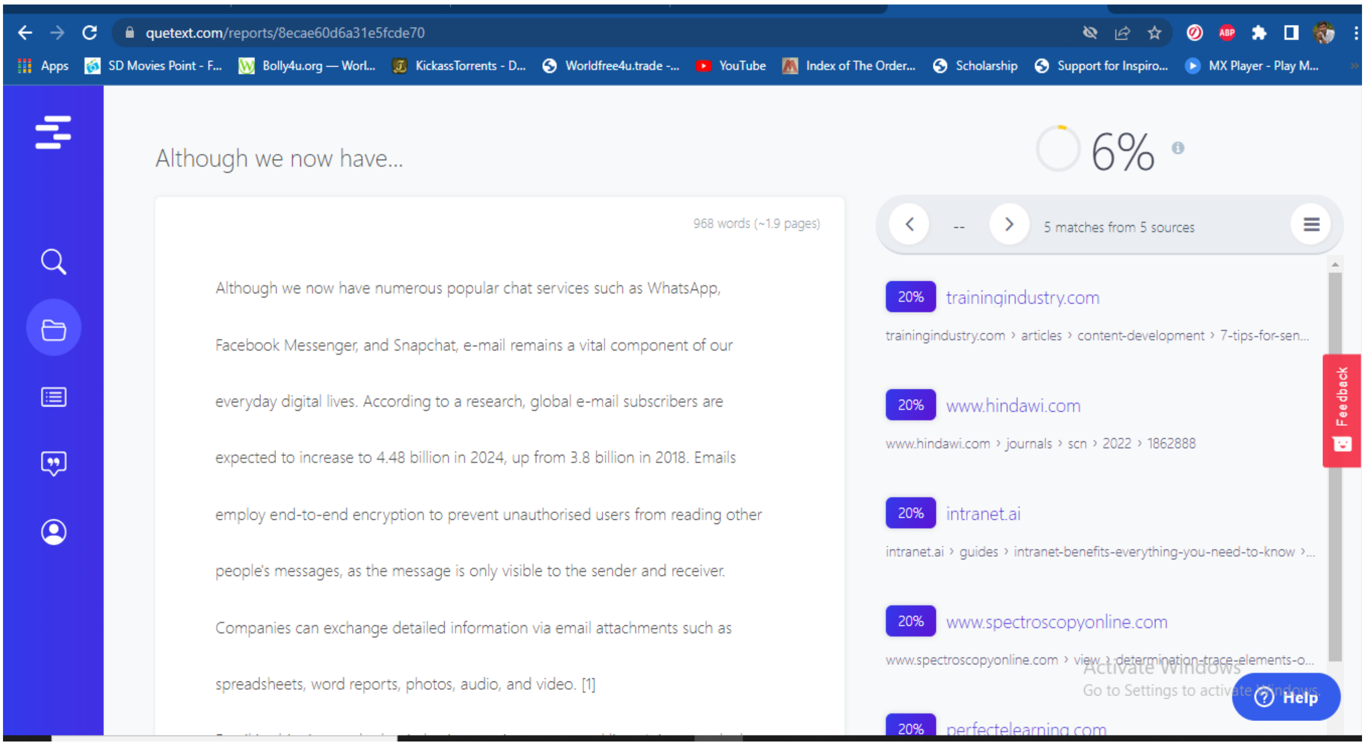
Designing Spam Mail Predictive System

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

# 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')
```

```
[0]
Spam mail
```

PLAGIARISM REPORT

Chapter 1 & 2:



Although we now have...

6%

968 words (~1.9 pages)

Although we now have numerous popular chat services such as WhatsApp, Facebook Messenger, and Snapchat, e-mail remains a vital component of our everyday digital lives. According to a research, global e-mail subscribers are expected to increase to 4.48 billion in 2024, up from 3.8 billion in 2018. Emails employ end-to-end encryption to prevent unauthorised users from reading other people's messages, as the message is only visible to the sender and receiver. Companies can exchange detailed information via email attachments such as spreadsheets, word reports, photos, audio, and video. [1]

-- 5 matches from 5 sources

20% trainingindustry.com
trainingindustry.com › articles › content-development › 7-tips-for-sen...

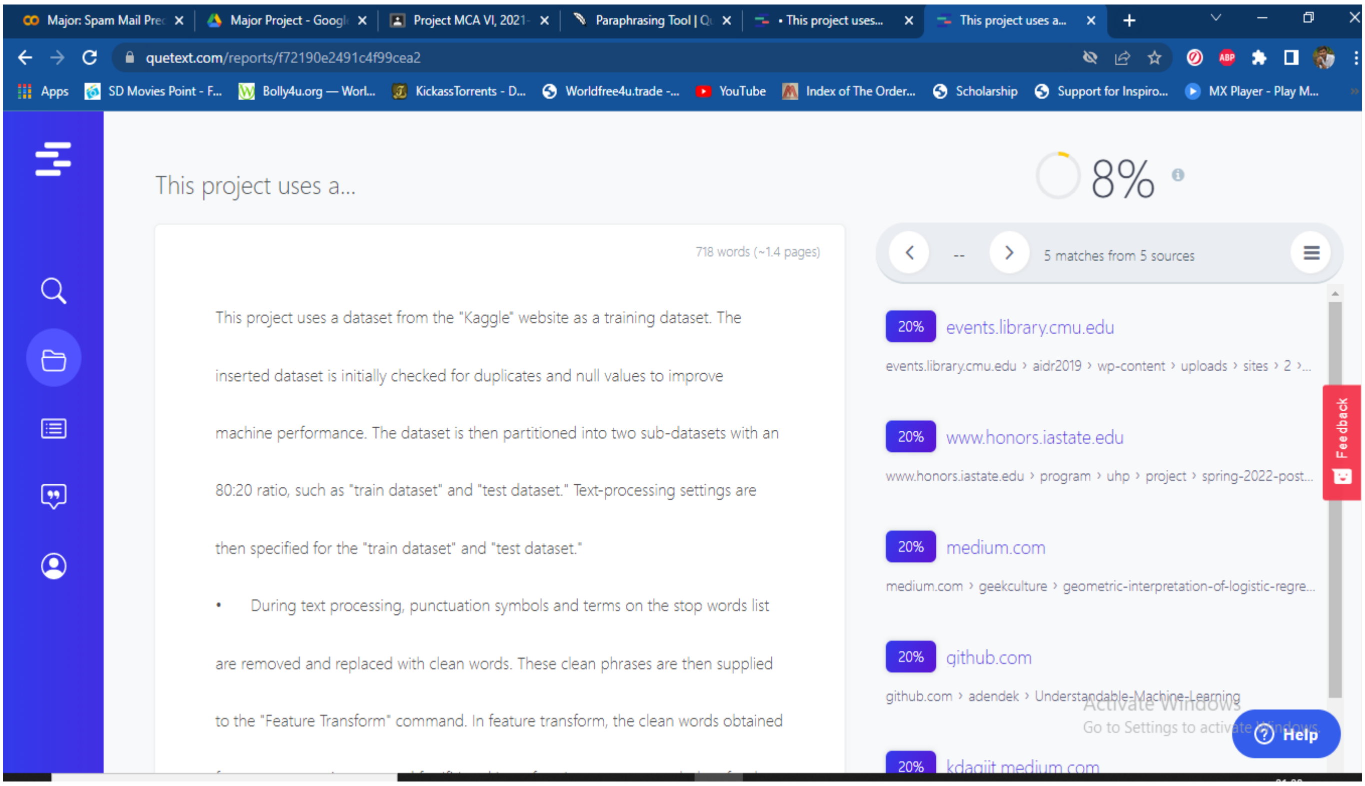
20% www.hindawi.com
www.hindawi.com › journals › scn › 2022 › 1862888

20% intranet.ai
intranet.ai › guides › intranet-benefits-everything-you-need-to-know ›...

20% www.spectroscopyonline.com
www.spectroscopyonline.com › view › determination-trace-elements-o...

20% perfectelearning.com

Chapter 3:



This project uses a...

8%

718 words (~1.4 pages)

This project uses a dataset from the "Kaggle" website as a training dataset. The inserted dataset is initially checked for duplicates and null values to improve machine performance. The dataset is then partitioned into two sub-datasets with an 80:20 ratio, such as "train dataset" and "test dataset." Text-processing settings are then specified for the "train dataset" and "test dataset."

- During text processing, punctuation symbols and terms on the stop words list are removed and replaced with clean words. These clean phrases are then supplied to the "Feature Transform" command. In feature transform, the clean words obtained

-- 5 matches from 5 sources

20% events.library.cmu.edu
events.library.cmu.edu › aidr2019 › wp-content › uploads › sites › 2 ›...

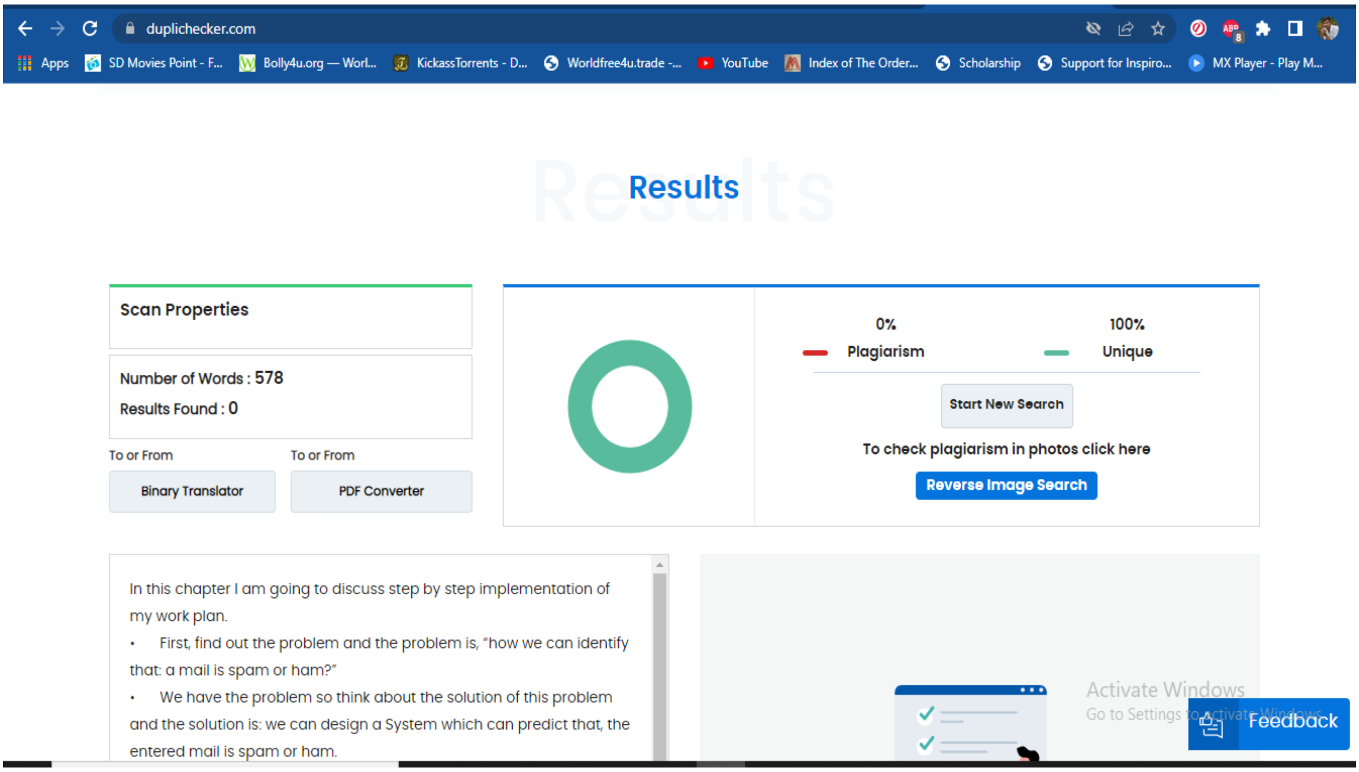
20% www.honors.iastate.edu
www.honors.iastate.edu › program › uhp › project › spring-2022-post...

20% medium.com
medium.com › geekculture › geometric-interpretation-of-logistic-regre...

20% github.com
github.com › adendek › Understandable-Machine-Learning

20% kdagiit.medium.com

Chapter 4 & 5:



Results

Scan Properties

Number of Words : 578

Results Found : 0

0% Plagiarism

100% Unique

In this chapter I am going to discuss step by step implementation of my work plan.

- First, find out the problem and the problem is, "how we can identify that: a mail is spam or ham?"
- We have the problem so think about the solution of this problem and the solution is: we can design a System which can predict that, the entered mail is spam or ham.