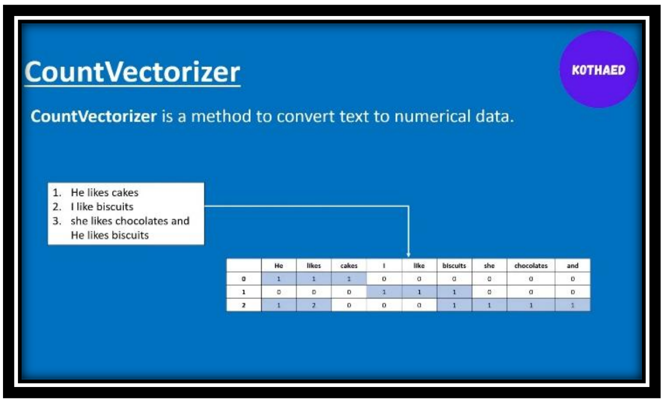
Project on Fundamentals of Machine Learning

Problem statement: To predict whether an incoming E-Mail message is spam or not and alert the user.

Theory:

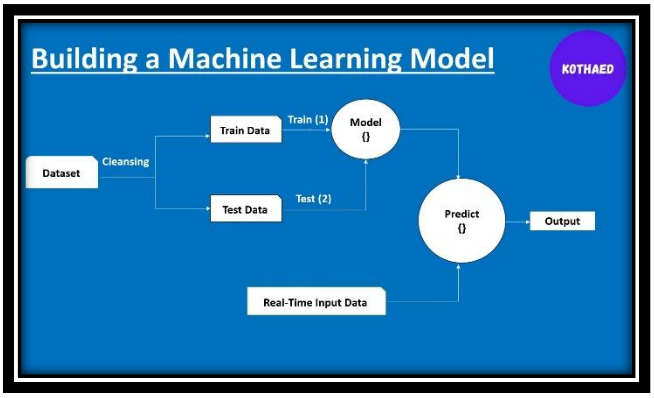


CountVectorizer

CountVectorizer is a method to convert text to numerical data.

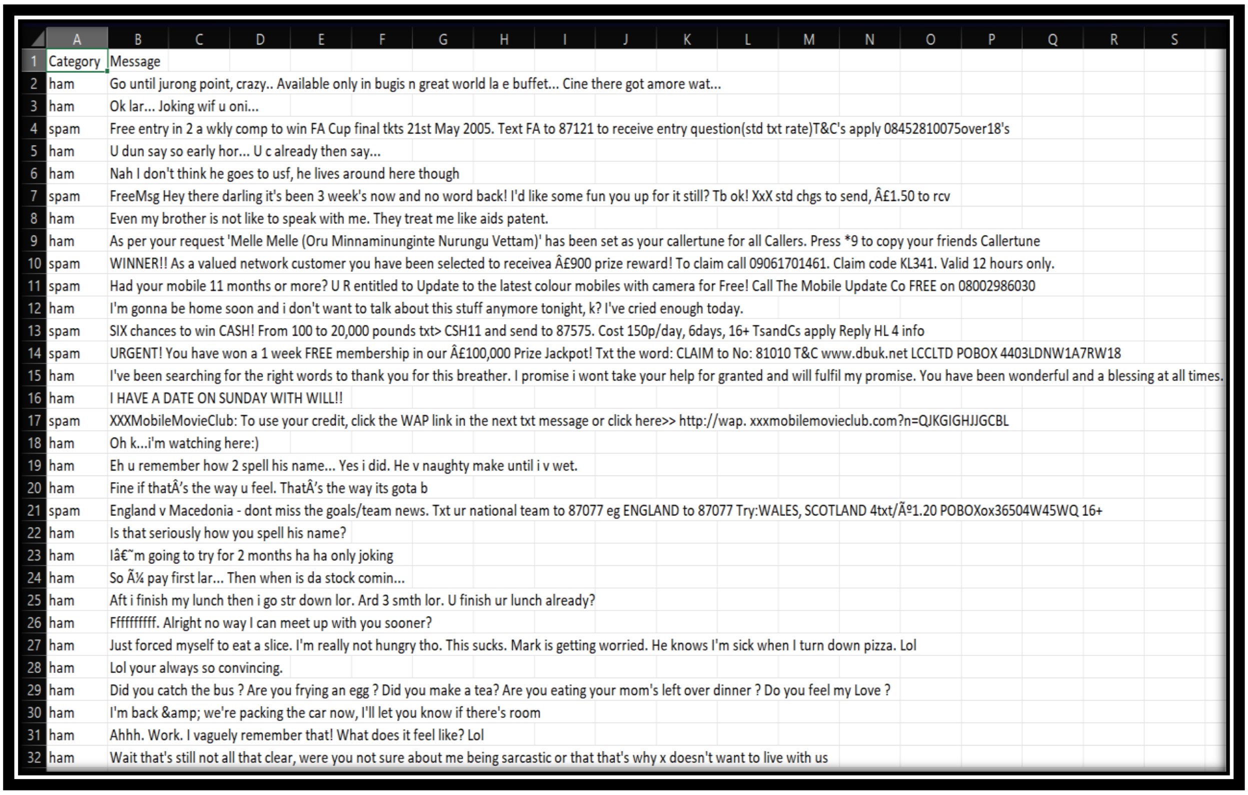
1. He likes cakes
2. I like biscuits
3. she likes chocolates and He likes biscuits

	He	likes	cakes	I	like	biscuits	she	chocolates	and
0	1	1	1	0	0	0	0	0	0
1	0	0	0	1	1	1	0	0	0
2	1	2	0	0	0	1	1	1	1



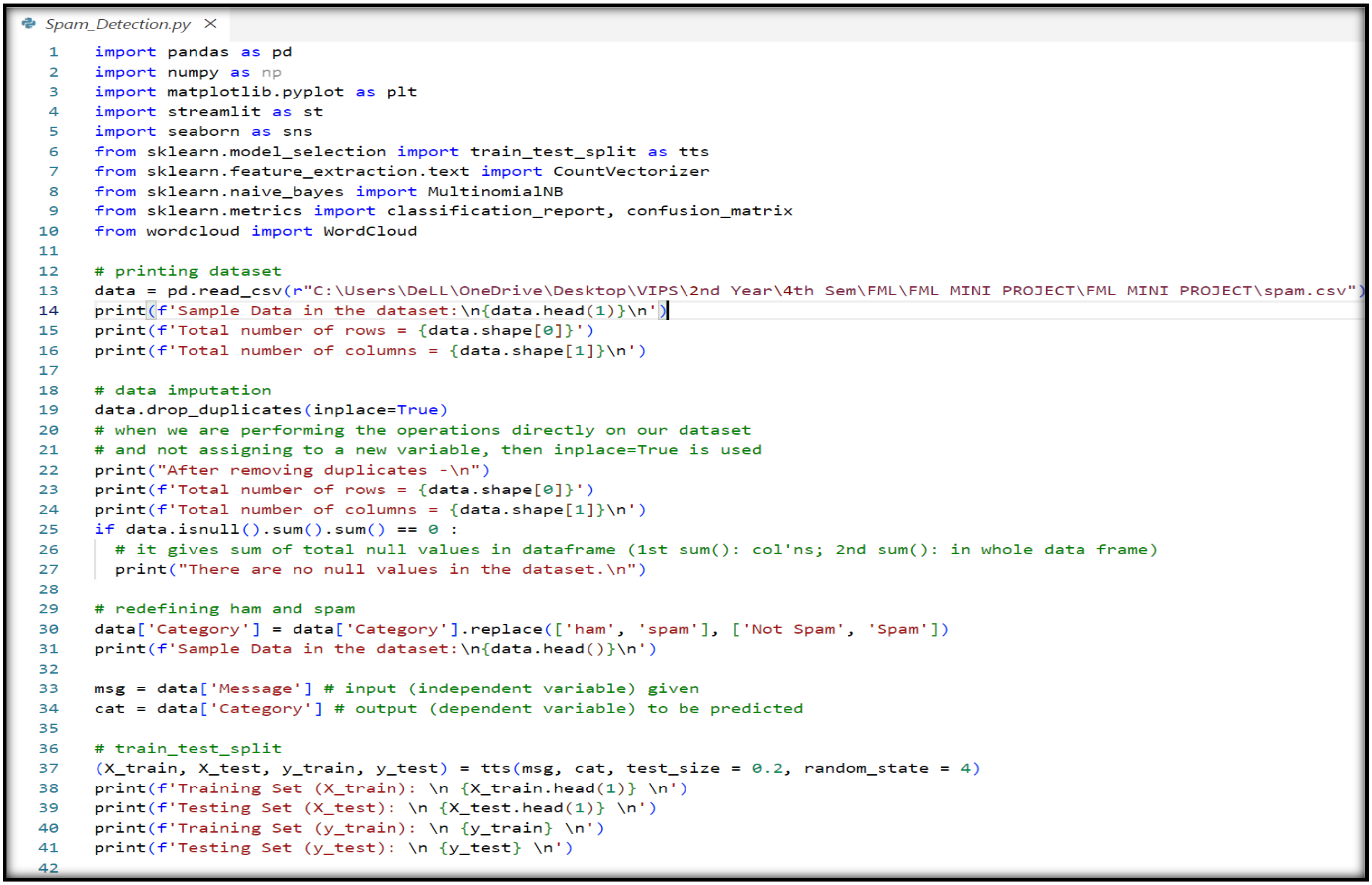
Building a Machine Learning Model

Dataset Used: -

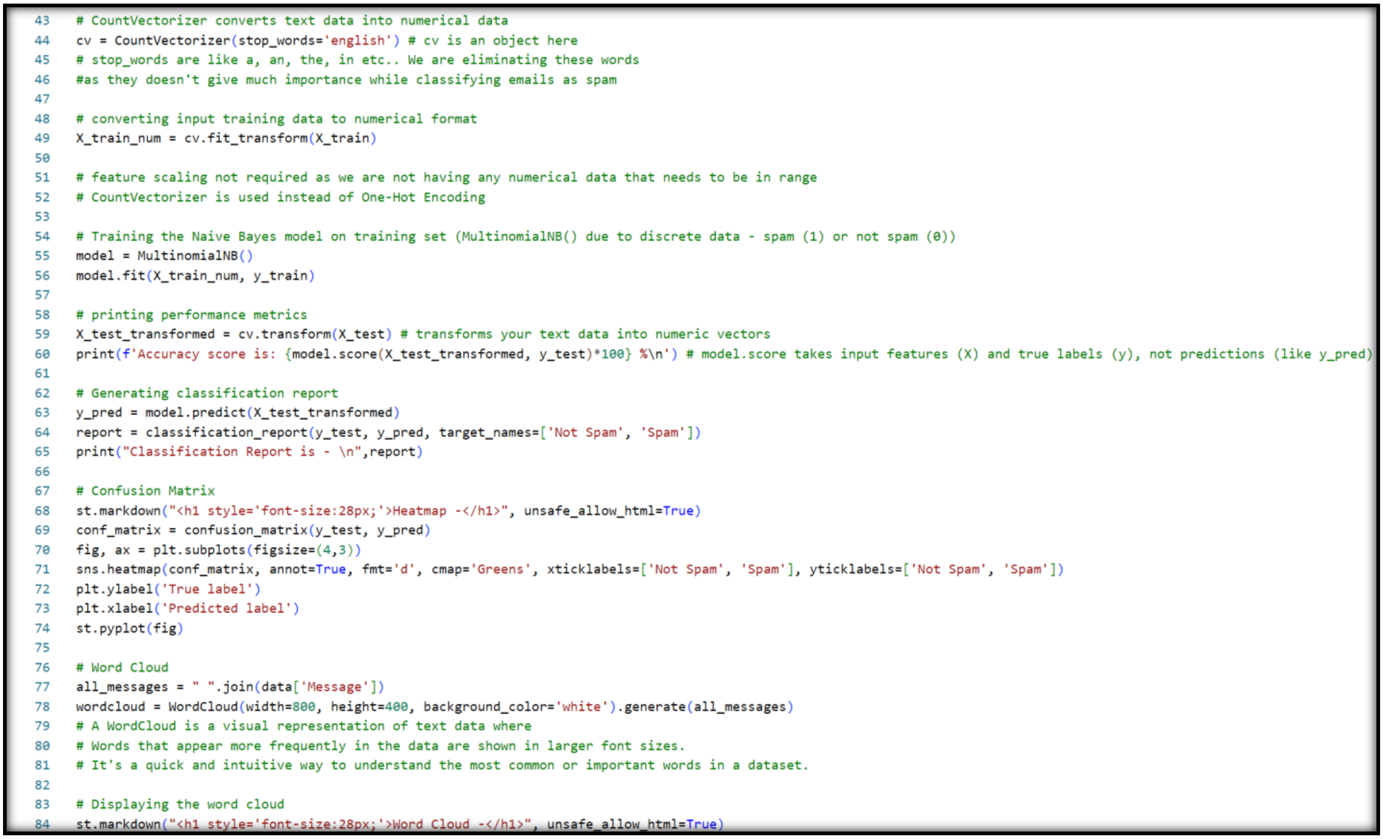


	A	B
1	Category	Message
2	ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
3	ham	Ok lar... Joking wif u oni...
4	spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
5	ham	U dun say so early hor... U c already then say...
6	ham	Nah I don't think he goes to usf, he lives around here though
7	spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, Â£1.50 to rcv
8	ham	Even my brother is not like to speak with me. They treat me like aids patent.
9	ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune
10	spam	WINNER!! As a valued network customer you have been selected to receivea Â£900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.
11	spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030
12	ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today.
13	spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info
14	spam	URGENT! You have won a 1 week FREE membership in our Â£100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18
15	ham	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.
16	ham	I HAVE A DATE ON SUNDAY WITH WILL!!
17	spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?n=QJKGIGHJJGCBL
18	ham	Oh k...i'm watching here:)
19	ham	Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
20	ham	Fine if thatÂ’s the way u feel. ThatÂ’s the way its gota b
21	spam	England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES, SCOTLAND 4txt/Ãº1.20 POBOXox36504W45WQ 16+
22	ham	Is that seriously how you spell his name?
23	ham	Iâ€˜m going to try for 2 months ha ha only joking
24	ham	So Ã¼ pay first lar... Then when is da stock comin...
25	ham	Aft i finish my lunch then i go str down lor. Ard 3 smth lor. U finish ur lunch already?
26	ham	Fffffffff. Alright no way I can meet up with you sooner?
27	ham	Just forced myself to eat a slice. I'm really not hungry tho. This sucks. Mark is getting worried. He knows I'm sick when I turn down pizza. Lol
28	ham	Lol your always so convincing.
29	ham	Did you catch the bus ? Are you frying an egg ? Did you make a tea? Are you eating your mom's left over dinner ? Do you feel my Love ?
30	ham	I'm back &amp; we're packing the car now, I'll let you know if there's room
31	ham	Ahhh. Work. I vaguely remember that! What does it feel like? Lol
32	ham	Wait that's still not all that clear, were you not sure about me being sarcastic or that that's why x doesn't want to live with us

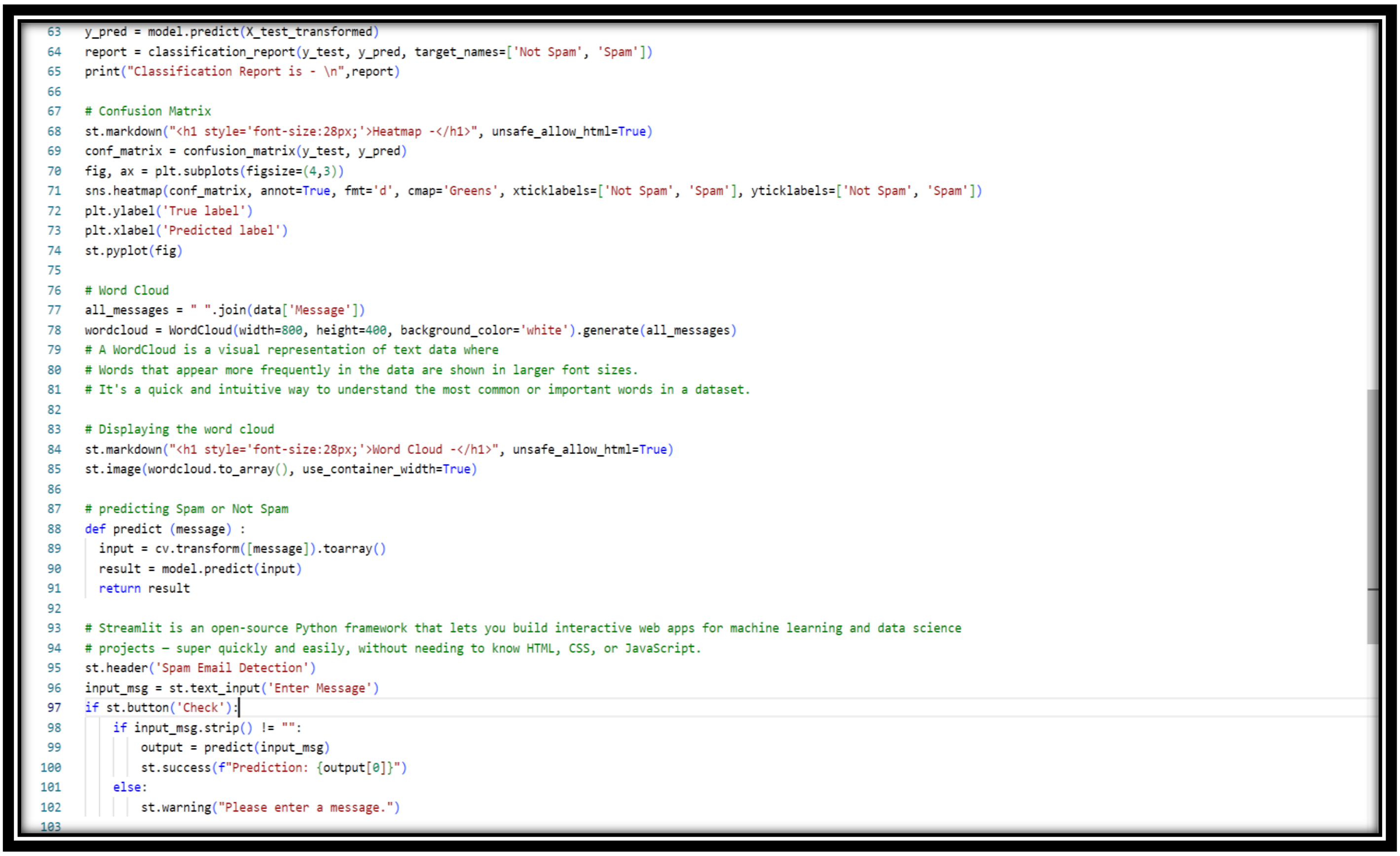
Source Code:



```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import streamlit as st
import seaborn as sns
from sklearn.model_selection import train_test_split as tts
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import classification_report, confusion_matrix
from wordcloud import WordCloud

# printing dataset
data = pd.read_csv(r"C:\Users\DeLL\OneDrive\Desktop\VIPS\2nd Year\4th Sem\FML\FML MINI PROJECT\FML MINI PROJECT\spam.csv")
print(f'Sample Data in the dataset:\n{data.head(1)}\n')
print(f'Total number of rows = {data.shape[0]}')
print(f'Total number of columns = {data.shape[1]}\n')

# data imputation
data.drop_duplicates(inplace=True)
# when we are performing the operations directly on our dataset
# and not assigning to a new variable, then inplace=True is used
print("After removing duplicates -\n")
print(f'Total number of rows = {data.shape[0]}')
print(f'Total number of columns = {data.shape[1]}\n')
if data.isnull().sum().sum() == 0 :
    # it gives sum of total null values in dataframe (1st sum(): col'ns; 2nd sum(): in whole data frame)
    print("There are no null values in the dataset.\n")

# redefining ham and spam
data['Category'] = data['Category'].replace(['ham', 'spam'], ['Not Spam', 'Spam'])
print(f'Sample Data in the dataset:\n{data.head()}\n')

msg = data['Message'] # input (independent variable) given
cat = data['Category'] # output (dependent variable) to be predicted

# train_test_split
(X_train, X_test, y_train, y_test) = tts(msg, cat, test_size = 0.2, random_state = 4)
print(f'Training Set (X_train): \n {X_train.head(1)} \n')
print(f'Testing Set (X_test): \n {X_test.head(1)} \n')
print(f'Training Set (y_train): \n {y_train} \n')
print(f'Testing Set (y_test): \n {y_test} \n')

```

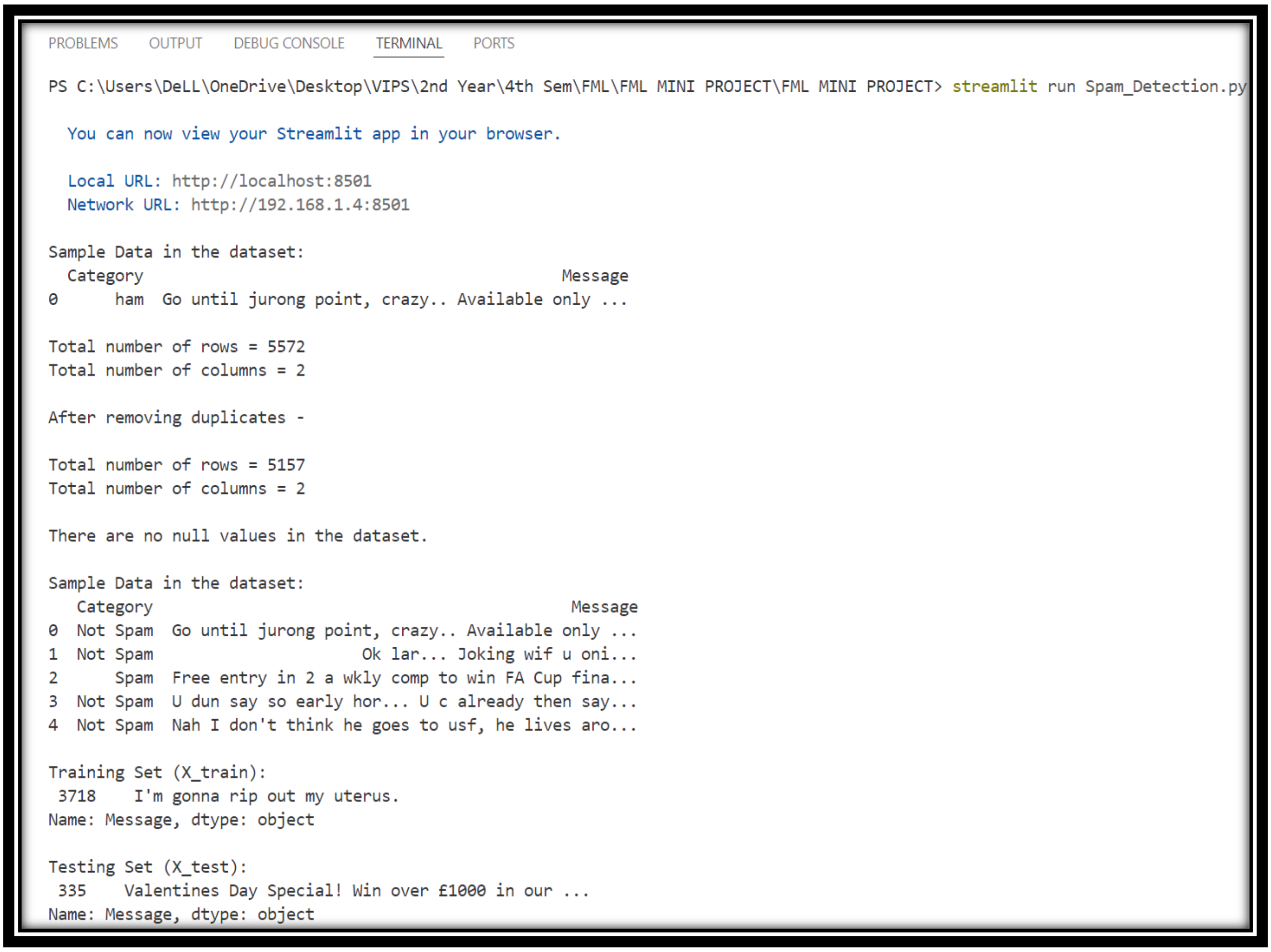


```python
# CountVectorizer converts text data into numerical data
cv = CountVectorizer(stop_words='english') # cv is an object here
# stop_words are like a, an, the, in etc.. We are eliminating these words
#as they doesn't give much importance while classifying emails as spam

# converting input training data to numerical format
X_train_num = cv.fit_transform(X_train)

# feature scaling not required as we are not having any numerical data that needs to be in range
# CountVectorizer is used instead of One-Hot Encoding

# Training the Naive Bayes model on training set (MultinomialNB() due to discrete data - spam (1) or not spam (0))
model = MultinomialNB()
model.fit(X_train_num, y_train)

# printing performance metrics
X_test_transformed = cv.transform(X_test) # transforms your text data into numeric vectors
print(f'Accuracy score is: {model.score(X_test_transformed, y_test)*100} %\n') # model.score takes input features (X) and true labels (y), not predictions (like y_pred)

# Generating classification report
y_pred = model.predict(X_test_transformed)
report = classification_report(y_test, y_pred, target_names=['Not Spam', 'Spam'])
print("Classification Report is - \n",report)

# Confusion Matrix
st.markdown("<h1 style='font-size:28px;'>Heatmap -</h1>", unsafe_allow_html=True)
conf_matrix = confusion_matrix(y_test, y_pred)
fig, ax = plt.subplots(figsize=(4,3))
sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Greens', xticklabels=['Not Spam', 'Spam'], yticklabels=['Not Spam', 'Spam'])
plt.ylabel('True label')
plt.xlabel('Predicted label')
st.pyplot(fig)

# Word Cloud
all_messages = " ".join(data['Message'])
wordcloud = WordCloud(width=800, height=400, background_color='white').generate(all_messages)
# A WordCloud is a visual representation of text data where
# Words that appear more frequently in the data are shown in larger font sizes.
# It's a quick and intuitive way to understand the most common or important words in a dataset.

# Displaying the word cloud
st.markdown("<h1 style='font-size:28px;'>Word Cloud -</h1>", unsafe_allow_html=True)
```

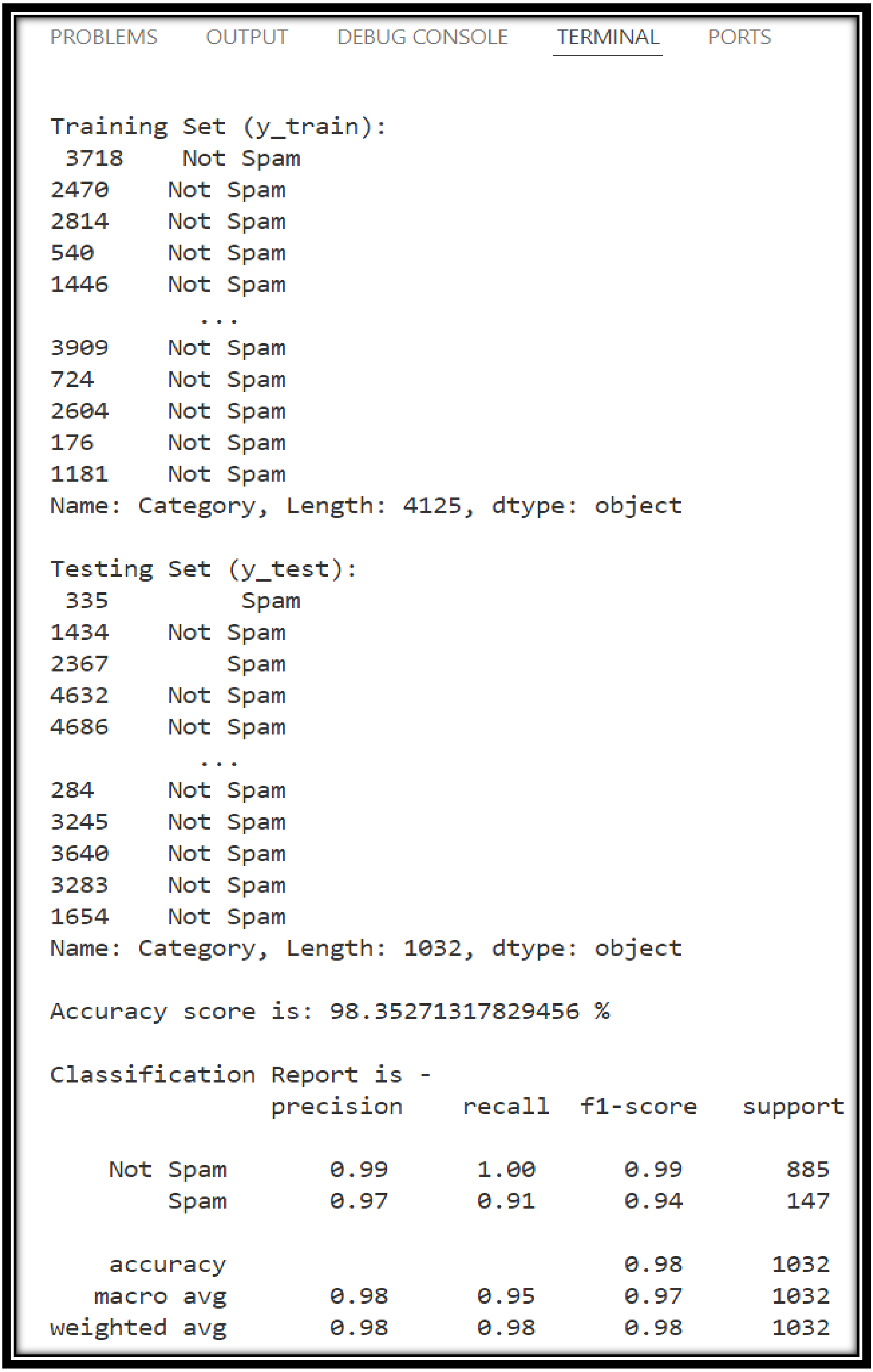


```python
y_pred = model.predict(X_test_transformed)
report = classification_report(y_test, y_pred, target_names=['Not Spam', 'Spam'])
print("Classification Report is - \n",report)

# Confusion Matrix
st.markdown("<h1 style='font-size:28px;'>Heatmap -</h1>", unsafe_allow_html=True)
conf_matrix = confusion_matrix(y_test, y_pred)
fig, ax = plt.subplots(figsize=(4,3))
sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Greens', xticklabels=['Not Spam', 'Spam'], yticklabels=['Not Spam', 'Spam'])
plt.ylabel('True label')
plt.xlabel('Predicted label')
st.pyplot(fig)

# Word Cloud
all_messages = " ".join(data['Message'])
wordcloud = WordCloud(width=800, height=400, background_color='white').generate(all_messages)
# A WordCloud is a visual representation of text data where
# Words that appear more frequently in the data are shown in larger font sizes.
# It's a quick and intuitive way to understand the most common or important words in a dataset.

# Displaying the word cloud
st.markdown("<h1 style='font-size:28px;'>Word Cloud -</h1>", unsafe_allow_html=True)
st.image(wordcloud.to_array(), use_container_width=True)

# predicting Spam or Not Spam
def predict (message) :
  input = cv.transform([message]).toarray()
  result = model.predict(input)
  return result

# Streamlit is an open-source Python framework that lets you build interactive web apps for machine learning and data science
# projects — super quickly and easily, without needing to know HTML, CSS, or JavaScript.
st.header('Spam Email Detection')
input_msg = st.text_input('Enter Message')
if st.button('Check'):
    if input_msg.strip() != "":
        output = predict(input_msg)
        st.success(f"Prediction: {output[0]}")
    else:
        st.warning("Please enter a message.")
```

Outputs: -

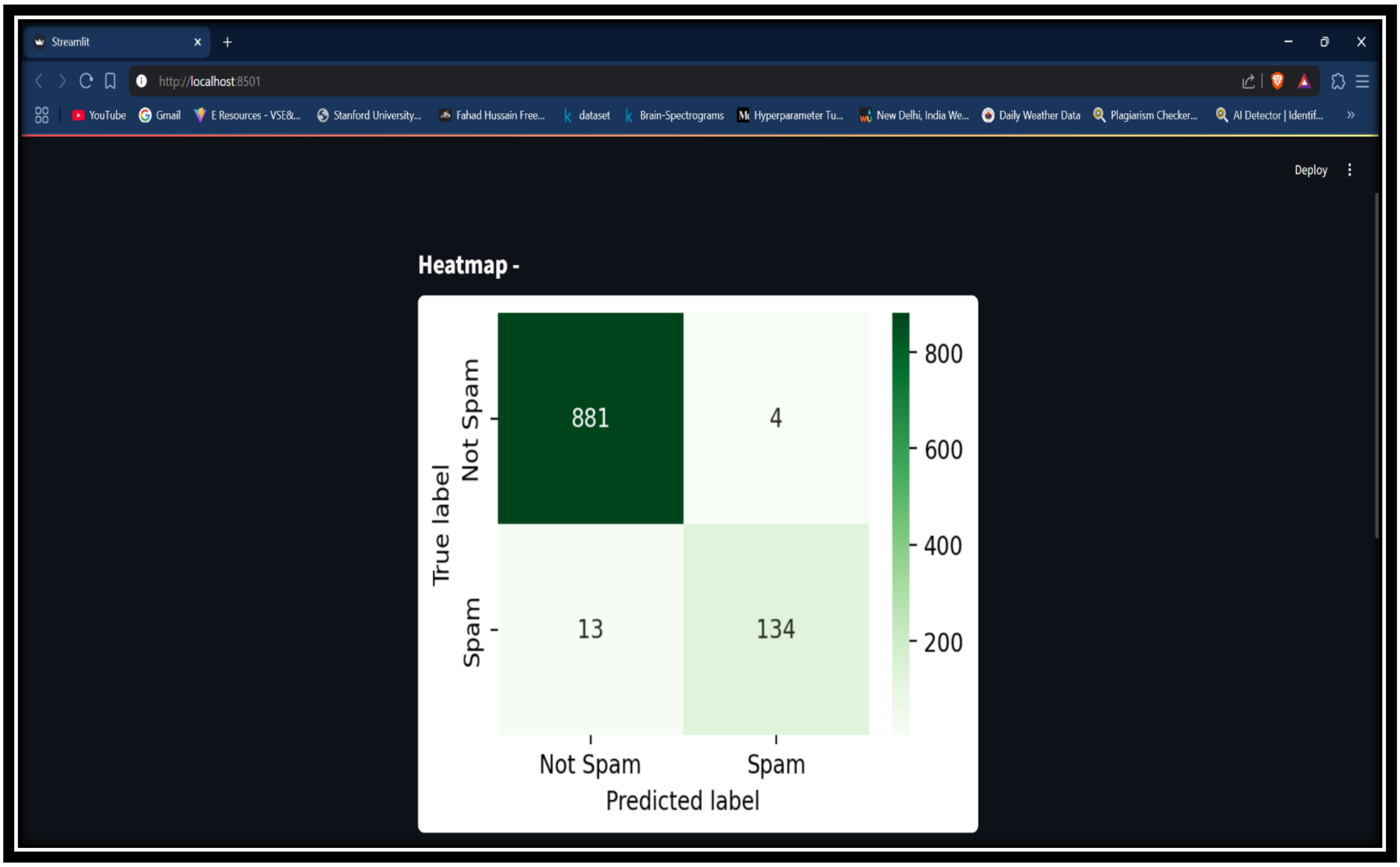


```
PS C:\Users\DeLL\OneDrive\Desktop\VIPS\2nd Year\4th Sem\FML\FML MINI PROJECT\FML MINI PROJECT> streamlit run Spam_Detection.py

  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8501
  Network URL: http://192.168.1.4:8501

Sample Data in the dataset:
  Category                                            Message
0      ham  Go until jurong point, crazy.. Available only ...

Total number of rows = 5572
Total number of columns = 2

After removing duplicates -

Total number of rows = 5157
Total number of columns = 2

There are no null values in the dataset.

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

Training Set (X_train):
 3718    I'm gonna rip out my uterus.
Name: Message, dtype: object

Testing Set (X_test):
 335    Valentines Day Special! Win over £1000 in our ...
Name: Message, dtype: object
```

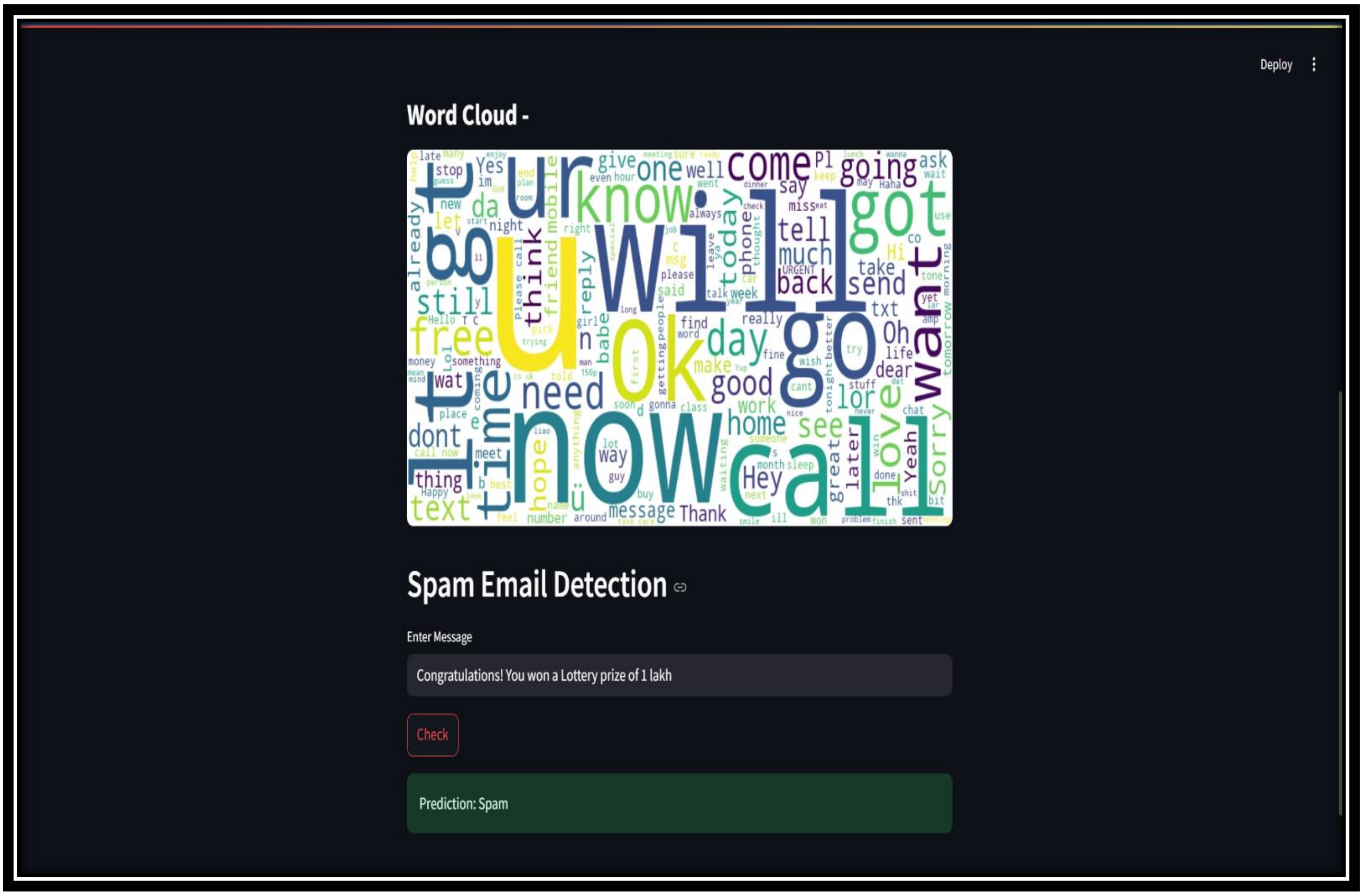


```
Training Set (y_train):
 3718    Not Spam
2470    Not Spam
2814    Not Spam
540     Not Spam
1446    Not Spam
          ...
3909    Not Spam
724     Not Spam
2604    Not Spam
176     Not Spam
1181    Not Spam
Name: Category, Length: 4125, dtype: object

Testing Set (y_test):
 335         Spam
1434    Not Spam
2367        Spam
4632    Not Spam
4686    Not Spam
          ...
284     Not Spam
3245    Not Spam
3640    Not Spam
3283    Not Spam
1654    Not Spam
Name: Category, Length: 1032, dtype: object

Accuracy score is: 98.35271317829456 %

Classification Report is -
               precision    recall  f1-score   support

    Not Spam       0.99      1.00      0.99       885
        Spam       0.97      0.91      0.94       147

    accuracy                           0.98      1032
   macro avg       0.98      0.95      0.97      1032
weighted avg       0.98      0.98      0.98      1032
```

Output hosted on web using Streamlit: -



Heatmap -



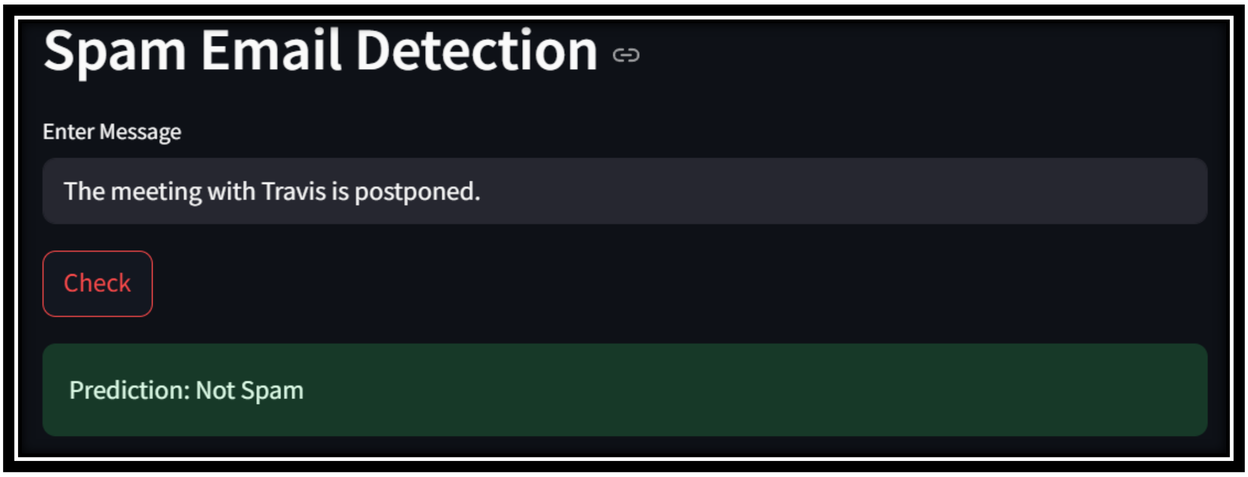
Word Cloud -

Spam Email Detection

Enter Message

Congratulations! You won a Lottery prize of 1 lakh

Prediction: Spam



Spam Email Detection

Enter Message

The meeting with Travis is postponed.

Prediction: Not Spam

Learning Outcome: