Building a Smarter AI-Powered Spam Classifier

Project Overview:

In this part you will need to understand the problem statement and create a document on what have you understood and how will you proceed ahead with solving the problem. Please think on a design and present in form of a document.

Problem Definition:

The problem is to build an AI-powered spam classifier that can accurately distinguish between spam and non-spam messages in emails or text messages. The goal is to reduce the number of false positives (classifying legitimate messages as spam) and false negatives (missing actual spam messages) while achieving a high level of accuracy.

Design Thinking:

1.Data Collection: We will need a dataset containing labeled examples of spam and nonspam messages. We can use a Kaggle dataset for this purpose.

2.Data Preprocessing: The text data needs to be cleaned and preprocessed. This involves removing special characters, converting text to lowercase, and tokenizing the text into individual words.

3.Feature Extraction: We will convert the tokenized words into numerical features using techniques like TF-IDF (Term Frequency-Inverse Document Frequency).

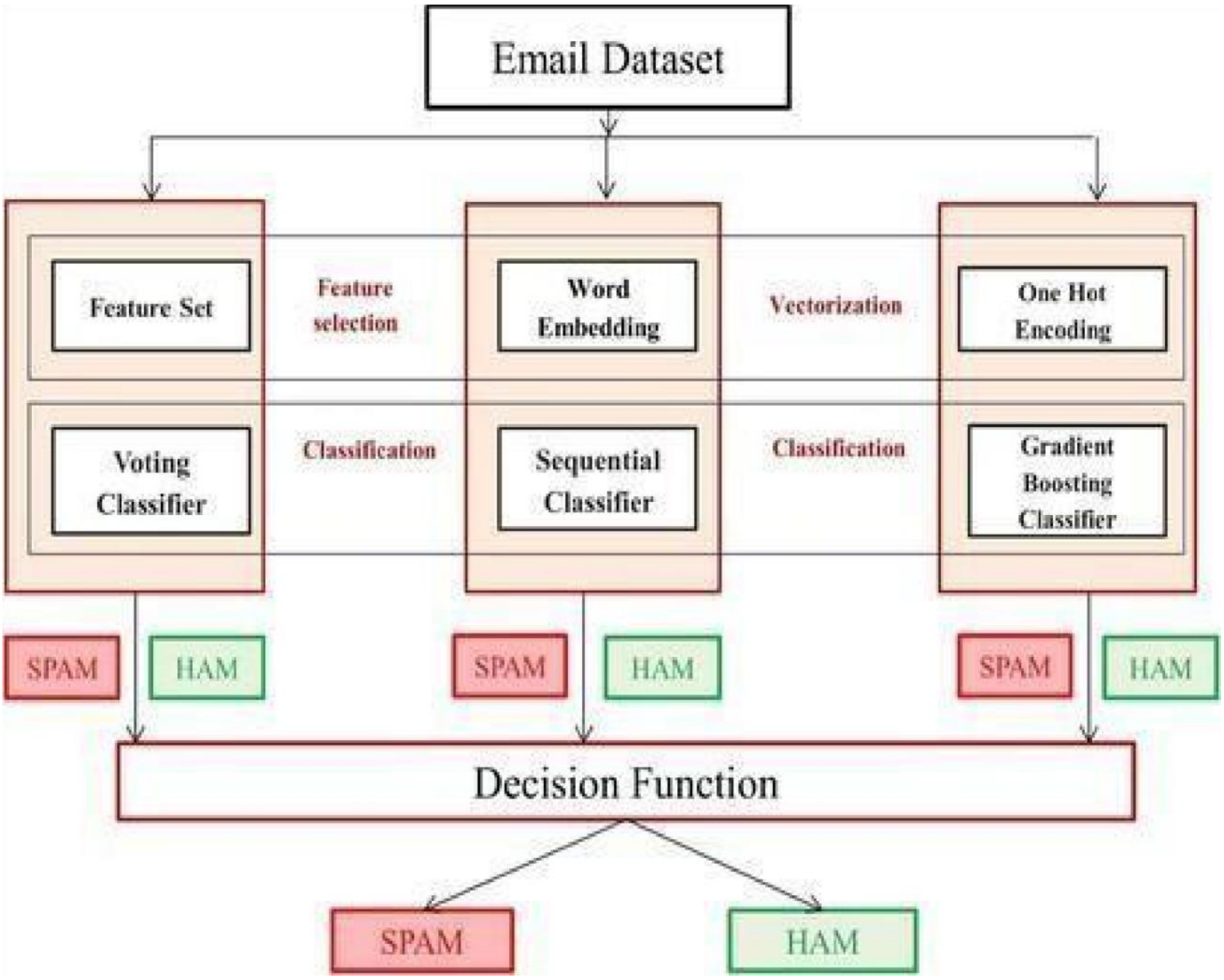
4.Model Selection: We can experiment with various machine learning algorithms such as Naive Bayes, Support Vector Machines, and more advanced techniques like deep learning using neural networks.

5.Evaluation: We will measure the model's performance using metrics like accuracy, precision, recall, and F1-score.

6**.Iterative Improvement:** We will fine-tune the model and experiment with hyperparameters to improve its accuracy.

Dataset Link: https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset

ARCHITECTURE FOR SPAM CLASSIFIER



1. In this architecture we divide mail into two parts as SPAM and HAM.
2. We are using three types of classifiers to classify the mails following as:

- ○ Voting Classifer
- ○ Sequential Classifer
- ○ Gradient Bossting Classifier

3. **These three classifers are used to filter messages, if spam or not sapm.**

REQUIREMENTS:

Hardware: PC with Core-I7.
Software : Windows 11, SPYDER.

PHASE4: In this phase, We use various algorithm for spam classification.

Algorithms:

➢ Multi-layer perceptron

➢ K-Nearest Neighbour

➢ Decision Tree

➢ Support Vector Machine

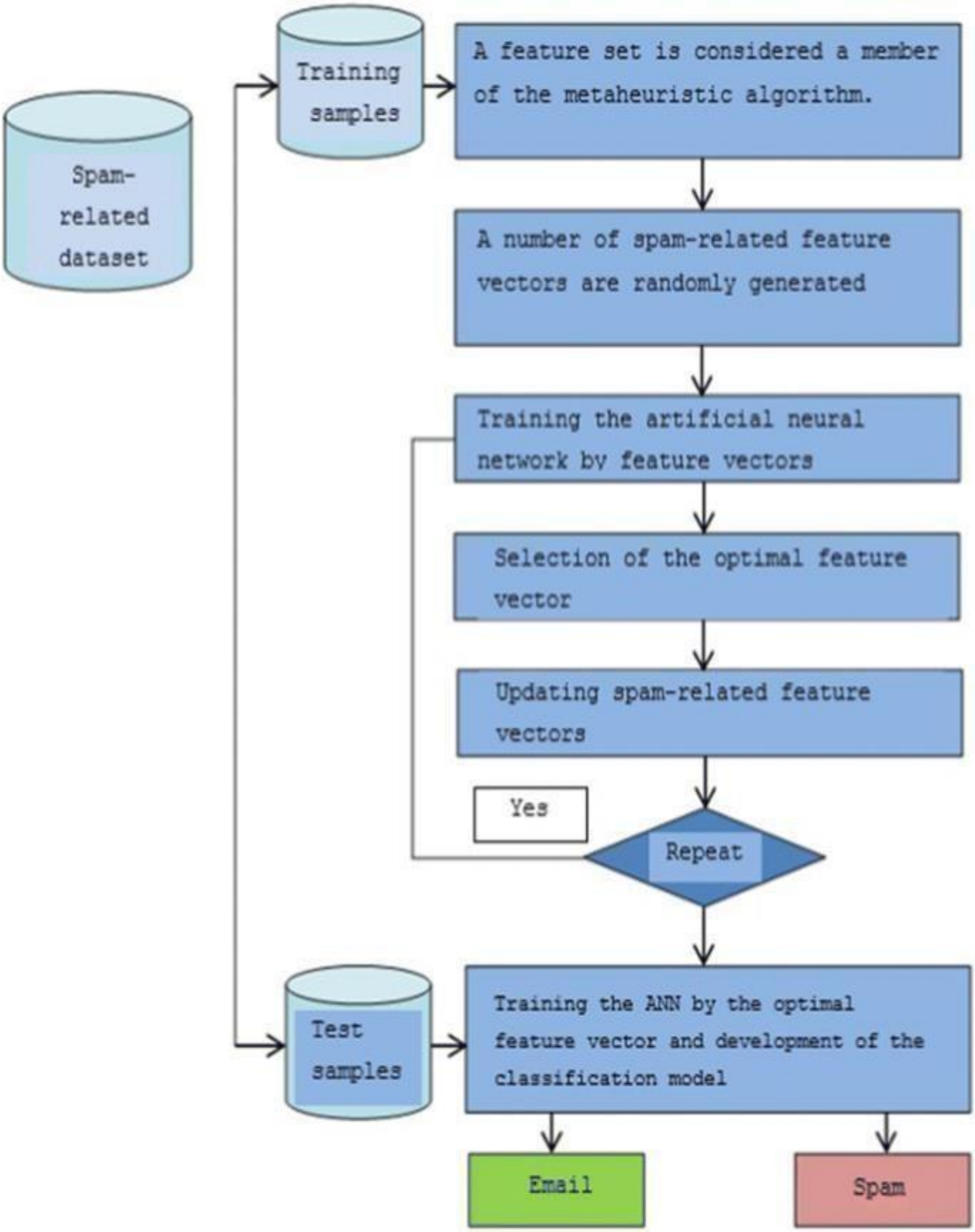
1. Multi-Layer Perceptron:

Abstract:

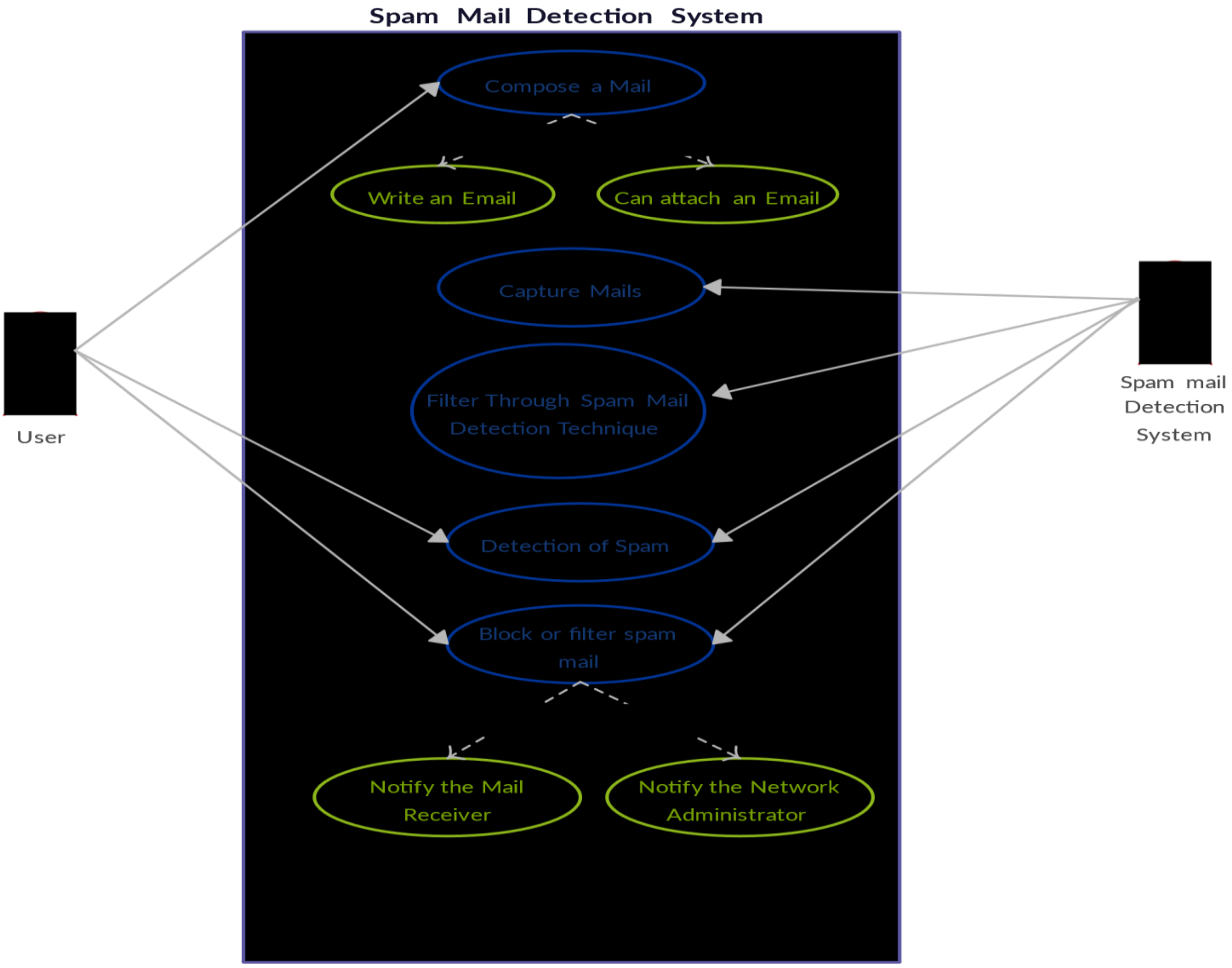
Spam email is a kind of junk email. Recent years, with the rapid growth of the internet users, especially emails users, the spam emails have been regarded as a severe problem. there are many classification methods that can be used to detect spam, Naïve Bayesian and Decision tree for example. These methods all gains good performance in most of cases, but the true positive rate and the false positive rate of them are not good enough. In this paper we designed a neural network based spam classification algorithm to filter spam. Our model is a classical multi-player perceptron composed of two hidden layer, one input layer and one output layer. By applying different threshold to plot the roc curve, we demonstrate that our method outperforms most of existed method. We also demonstrate that ensemble learning will boost the whole method.

Keywords: Multi-layer perceptron, spam, ADAM, ROC

FLOWCHART


Spam-related dataset
Training samples
A feature set is considered a member of the metaheuristic algorithm.
A number of spam-related feature vectors are randomly generated
Training the artificial neural network by feature vectors
Selection of the optimal feature vector
Updating spam-related feature vectors
Yes
Repeat
Test samples
Training the ANN by the optimal feature vector and development of the classification model
Email
Spam

USER DIAGRAM FOR SPAM CLASSIFIER

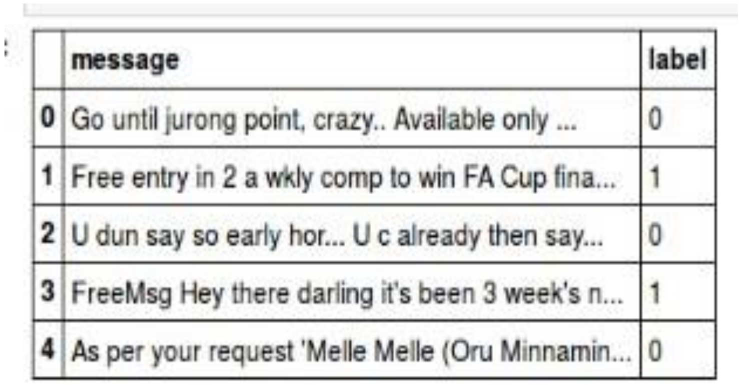

Spam Mail Detection System
Compose a Mail
Write an Email
Can attach an Email
Capture Mails
Filter Through Spam Mail Detection Technique
Detection of Spam
Block or filter spam mail
Notify the Mail Receiver
Notify the Network Administrator
User
Spam mail Detection System

1. Load and simplify the dataset.
2. Explore the dataset.
3. Handle the imbalance datasets.
4. Split the dataset.
5. Classify the dataset using algorithms like Random Forest, Naïve Bayes Etc.,
6. Finally find the messages spam or not spam.

PROGRAM:

```
import pandas as pd from sklearn.neural_network
 import MLPClassifier from sklearn.model_selection
 import train_test_split from sklearn.metrics accuracy_score
 # Load the dataset data = pd.read_csv('spam.csv')
# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(data['text'], data['label'],
random_state=0)
# Create an MLP classifier with two hidden layers of 16 and 8 neurons each
 mlp = MLPClassifier(hidden_layer_sizes=(16, 8), max_item=1000)
 # Train the classifier on the training set
 mlp.fit(X_train, y_train)
# Predict the labels of the test set
y_pred = mlp.predict(X_test)
# Print the accuracy of the classifier on the test set
print('Accuracy:', accuracy_score(y_test, y_pred))
```

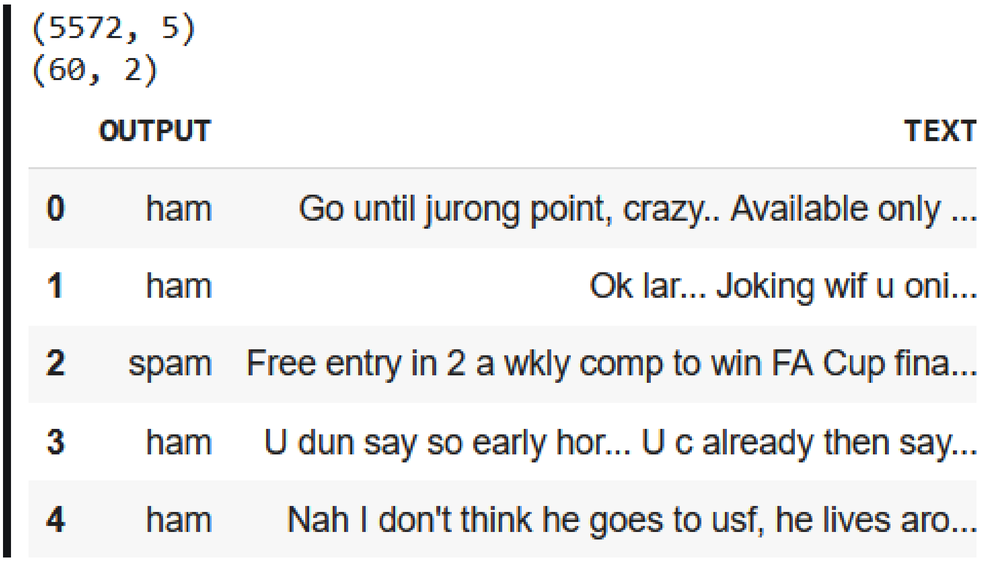
OUTPUT:



	message	label
0	Go until jurong point, crazy.. Available only ...	0
1	Free entry in 2 a wkly comp to win FA Cup fina...	1
2	U dun say so early hor... U c already then say...	0
3	FreeMsg Hey there darling it's been 3 week's n...	1
4	As per your request 'Melle Melle (Oru Minnamin...	0

```
# necessary libraries
import openai
import pandas as pd
import numpy as np
# libraries to develop and evaluate a machine learning model
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, accuracy_score
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, accuracy_score
from sklearn.metrics import confusion_matrix
# replace "YOUR API KEY" with your generated API key
openai.api_key = "YOUR API KEY"
# while loading the csv, we ignore any encoding errors and skip any bad line
df = pd.read_csv('spam.csv', encoding_errors='ignore', on_bad_lines='skip')
print(df.shape)
# we have 3 columns with NULL values, to remove that we use the below line
df = df.dropna(axis=1)
# we are taking only the first 60 rows for developing the model
df = df.iloc[:60]
# rename the columns v1 and v2 to Output and Text respectively
df.rename(columns = {'v1':'OUTPUT', 'v2': 'TEXT'}, inplace = True)
print(df.shape)
df.head()
```

OUTPUT:



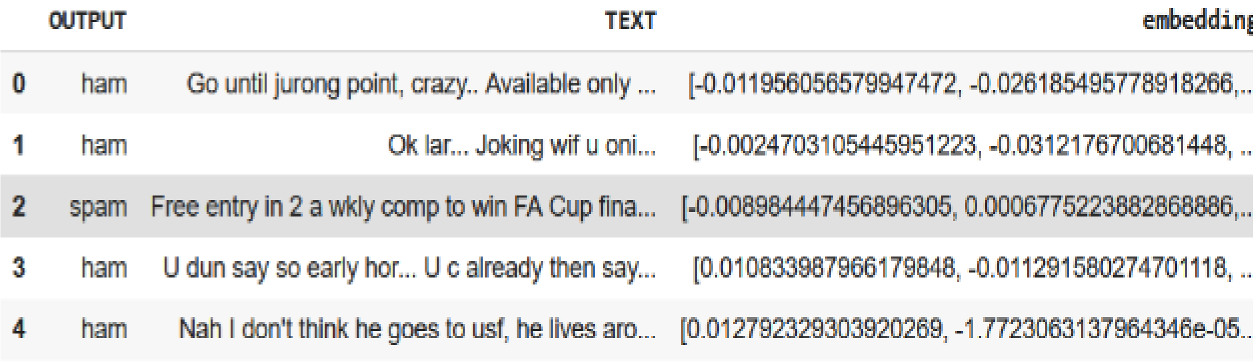
```
(5572, 5)
(60, 2)
```

	OUTPUT	TEXT
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...

```
# function to generate vector for a string
def get_embedding(text, model="text-embedding-ada-002"):
return openai.Embedding.create(input = , model=model)['data'][0]['embedding']

# applying the above funtion to generate vectors for all 60 text pieces
df["embedding"] = df.TEXT.apply(get_embedding).apply(np.array) # convert string to array
df.head()
```

OUTPUT:



	OUTPUT	TEXT	embedding
0	ham	Go until jurong point, crazy.. Available only ...	[-0.011956056579947472, -0.026185495778918266,..
1	ham	Ok lar... Joking wif u oni...	[-0.0024703105445951223, -0.0312176700681448, ..
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	[-0.008984447456896305, 0.000677522388286886,..
3	ham	U dun say so early hor... U c already then say...	[0.010833987966179848, -0.011291580274701118, ..
4	ham	Nah I don't think he goes to usf, he lives aro...	[0.012792329303920269, -1.7723063137964346e-05..

```
print("accuracy: ", np.round(accuracy_score(y_test, preds)*100,2), "%")
```

OUTPUT:

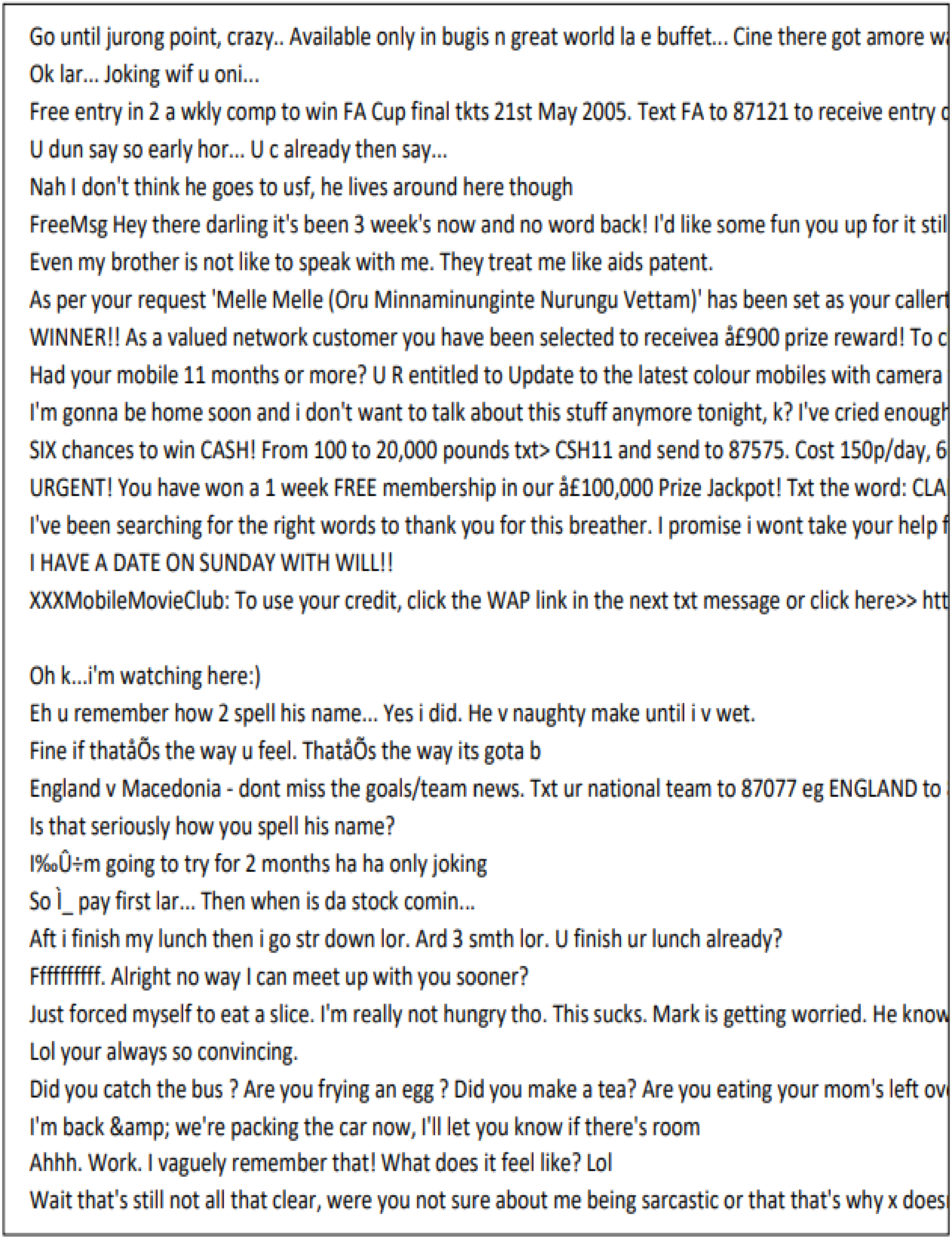
Accuracy: 83.33%

```
out = spam_classification("""Congratulations! You've Won a $1000 gift card Go to
                                  https://bit.ly to claim your reward.""")
print(out)
```

OUTPUT:

SPAM

DATASET



Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wa
Ok lar... Joking wif u oni...
Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry q
U dun say so early hor... U c already then say...
Nah I don't think he goes to usf, he lives around here though
FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it stil
Even my brother is not like to speak with me. They treat me like aids patent.
As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callert
WINNER!! As a valued network customer you have been selected to receivea å£900 prize reward! To c
Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera
I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough
SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6
URGENT! You have won a 1 week FREE membership in our å£100,000 Prize Jackpot! Txt the word: CLA
I've been searching for the right words to thank you for this breather. I promise i wont take your help f
I HAVE A DATE ON SUNDAY WITH WILL!!
XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> htt

Oh k...i'm watching here:)
Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
Fine if thatåÕs the way u feel. ThatåÕs the way its gota b
England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to
Is that seriously how you spell his name?
I‰Û÷m going to try for 2 months ha ha only joking
So Ì_ pay first lar... Then when is da stock comin...
Aft i finish my lunch then i go str down lor. Ard 3 smth lor. U finish ur lunch already?
Fffffffff. Alright no way I can meet up with you sooner?
Just forced myself to eat a slice. I'm really not hungry tho. This sucks. Mark is getting worried. He know
Lol your always so convincing.
Did you catch the bus ? Are you frying an egg ? Did you make a tea? Are you eating your mom's left ove
I'm back &amp; we're packing the car now, I'll let you know if there's room
Ahhh. Work. I vaguely remember that! What does it feel like? Lol
Wait that's still not all that clear, were you not sure about me being sarcastic or that that's why x does

CONCLUSION:

In conclusion, the developed SMS spam classifier successfully differentiates between spam and ham messages with high accuracy and precision. The **Random Forest Classifier** emerged as the best-performing providing consistent results on both training and test data.

Bayesian spam filtering is an incredibly powerful statistical technique—with acceptable computational complexity—for identifying spam messages.

Bayesian techniques address many weaknesses of other methodologies

- The entire message can be examined, not just special parts.
- All words are significant, not just special keywords or addresses.
- Updating is, in practice, infrequent (never more than one or two email per week through the training program; often none).
- So far, spam attacks on Bayesian filters have been relatively unsuccessful.
- When combined with other techniques, Bayesian filters can be a very strong component of an institution’s global spam system.

THANK YOU!!!