Classification Of Tweets/Messages Using Logistic Regression and Tf-idfVectoriser Enabled with Text-to-Speech

A PROJECT REPORT

Submitted by

SHIV NOLIYAN [Reg No: RA1711003010265]
NIKHIL KUMAR SINGH [Reg No: RA1711003010151]

Under the guidance of

M.SENTHIL RAJA
(Associate Professor, Department of Computer Science & Engineering)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

of

FACULTY OF ENGINEERING AND TECHNOLOGY

SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

S.R.M. Nagar, Kattankulathur, Kancheepuram District

MAY 2021

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Under Section 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

Certified that this project report titled “ **Classification Of Tweets / Messages Using Logistic Regression and Tf-idfVectoriser Enabled with Text-to-Speech**” is the bonafide work of **”SHIV NOLIYAN [Reg. No: RA1711003010265], NIKHIL KUMAR SINGH [Reg. No: RA1711003010151]”,** who carried out the projct work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form any other project report or dis-sertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Mr. M.Senthil Raja
GUIDE
Associate Professor
Dept. of Computer Science & Engineering

SIGNATURE

Dr. B. Amutha
HEAD OF THE DEPARTMENT
Dept. of Department of Computer Science and Engineering

Signature of the Internal Examiner

Signature of the External Examiner

Own Work Declaration

Department of Computer Science and Engineering

SRM Institute of Science & Technology

Own Work* Declaration Form

This sheet must be filled in (each box ticked to show that the condition has been met). It must be signed and dated along with your student registration number and included with all assignments you submit – work will not be marked unless this is done.

<u>To be completed by the student for all assessments</u>

Degree/ Course : B.Tech

Student Name : SHIV NOLIYAN ,

NIKHIL KUMAR SINGH

Registration Number : RA1711003010265

, RA1711003010151

Title of Work : Classification Of Tweets/Messages Using Logistic Regression and Tfidf-Vectoriser Enabled with Text-to-Speech

I / We hereby certify that this assessment compiles with the University's Rules and Regulations relating to Academic misconduct and plagiarism**, as listed in the University Website, Regulations, and the Education Committee guidelines.

I / We confirm that all the work contained in this assessment is my / our own except where indicated, and that I / We have met the following conditions:

- Clearly references / listed all sources as appropriate
- Referenced and put in inverted commas all quoted text (from books, web, etc)
- Given the sources of all pictures, data etc. that are not my own
- Not made any use of the report(s) or essay(s) of any other student(s) either past or present
- Acknowledged in appropriate places any help that I have received from others (e.g. fellow students, technicians, statisticians, external sources)
- Compiled with any other plagiarism criteria specified in the Course handbook / University website

I understand that any false claim for this work will be penalised in accordance with the University policies and regulations.

DECLARATION:

I am aware of and understand the University's policy on Academic misconduct and plagiarism and I certify that this assessment is my / our own work, except where indicated by referring, and that I have followed the good academic practices noted above.

If you are working in a group, please write your registration numbers and sign with the date for every student in your group.

SHIV NOLIYAN

RA1711003010265
SHIV NOLIYAN

Nikhil

RA1711003010151
NIKHIL KUMAR SINGH

ACKNOWLEDGEMENTS

We express our humble gratitude to **C. Muthamizhchelvan**, Vice Chancellor (I/C), SRM Institute of Science and Technology, for the facilities extended for the project work and his continued support.

We extend our sincere thanks to **Dr. Revathi Venkataraman**, Professor & Chairperson, School of Computing, SRM Institute of Science and Technology, for his invaluable support.
We wish to thank **Dr. B. Amutha**, Professor & Head, Department of Computer Science and Engineering, SRM Institute of Science and Technology, for her valuable suggestions and encouragement throughout the period of the project work.

We are extremely grateful to our Academic Advisor **Dr. P. Vishalakshi,** Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, for their great support at all the stages of project work. We would like to convey our thanks to our Panel Head, [**Dr.G. Manju],** [Designation], Department of Computer Science and Engineering, SRM Institute of Science and Technology, for his / her inputs during the project reviews.

We register our immeasurable thanks to our Faculty Advisor**, Ms. G. Malarselvi,** Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, for leading and helping us to complete our course.

Our inexpressible respect and thanks to my guide, **Mr. M.Senthil Raja,** Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, for providing me an opportunity to pursue my project under his/her mentorship. He / She provided me the freedom and support to explore the research topics of my interest. Her / His passion for solving the real problems and making a difference in the world has always been inspiring.

We sincerely thank staff and students of the Computer Science and Engineering Department, SRM Institute of Science and Technology, for their help during my research. Finally, we would like to thank my parents, our family members and our friends for their unconditional love, constant support and encouragement.

Shiv Noliyan

Nikhil Kumar Singh

ABSTRACT

This classification model's goal is to sort the massive influx of tweets/messages into ham/spam or authentic/fraud, i.e. to separate the useful from the unwanted. The TFIDFvectorizer technique is used in conjunction with Logistic Regression to improve the model's accuracy. The use of LR algorithm is for ham and spam in tweets/messages

Furthermore, utilising Word Cloud so to easy observed kind of tweets are commonly received bya customer or a corporation, i.e. either hamspam. Text-to-Speech is also included for a better user experience. The result is easy to watch & are more flexible to research the result.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

HTML	HyperText MarkUp Language
CSS	Cascading Style Sheet
IR	Information Retrieval
SQL	Structured Query Language
LR	Logistic Regression
UI	User Interface
JS	JavaScript
OTP	One-Time Password
UX	User Experience

LIST OF SYMBOLS

z	A real number
$h_\theta(x)$	Hypothesis function
$g(z)$	Sigmoid or logistic function
x	Constants

CHAPTER_1

INTRODUCTION

1.1 For what Purpose Tweet/Message Classifier System?

Inside Inovative world, day by day we used to get an exceptionally enormous number of tweets and messages fromdifferent organizations or individuals around the world. Once in a while, it is preposterous to expect to peruse all the messages got. Some of them might be critical while some are only for advancements, updates or datawhich is least significant. Manual message groupng is bound to get mistakes, will be tedious and furthermore will expand the expense to the organization. Along these lines,to manage this issue, a tweet/message characterization framework has been presented, which can order messages as per their sort and need.

A tweet/message classifier systemisa framework thatis intended to screen the continuous progressionof tweet/message inan association or in an personal record. Via naturally exploring and looking through the substance present in the body of the message and in anyconnections, the framework will group each tweet/message into different classes as requested by an organization or by a person. This spares time just as odds of mistakes and besides itdiminishes the representative expense to an organization.

1.2 Types of Tweets/Messages

There are various classes of messages like

- Ham
- Spam
- Promotions
- Forum
- Primary
- Ad.

1.3 What is WordCloud and Why is it Used?

Word cloud is a method of depicting the word as indicated by their recurrence. It shows the nature of the word according to freq-uecy of the word present in the collection.

1.3.1 Benefits of word cloud-

- Word clouds are straightforward andgive clearness.
- Word clouds are aground-breaking specialized apparatus.
- Word clouds are outwardly appealing thanplain information.

CHAPTER_2

LITERATURE SURVEY

“spammers tend to favor its use in spreading their commercial messages due to the high popularity of Twitter,. In the context of detecting twitter spams, behavioral analysis and different statical approaches were proposed. However, these techniques suffer from many limitations due to”[1]. constraints access the user’s listof followersorfollowees by ongoing change toTwitter’s streamingAPI which spammer’s creativitness in building diverse messages is done, and use of embeddedlinks and fresh accounts, & needfor analyzing diff. characteristic about user’s without theirconsent. compared to-our ontology-based approach which out-performs them by approx. 200% .our experiment conduct on real tweets data which illustrates that message to-message technique achieved low detection-rate ”[4].

Neural word embeddings have recently been used as less time-consuming representations than manual feature engineering. The majority of these word-embeddings model word syntactic information while ignoring sentiment context[5].. Close by this, a system that gives a yield that is customer express has been centered around. This ensures a dominating customer experience for every individual who uses the structure. “NLPtechniques stop-words removing, lemmatizing, stemming have been tried onboth algorithms to-inspect differences in-accuracy aswellas to find best method among those”. So in this review paper we focused on sentiment analysis of Twitter data.

2.1 Inference from the survey papers

From the overview deduction it is found that the precision can be gotten rather through consisting of multiple algorithms in a tweet arrangement framework. Be that as it may, by which includes numerous message classification algorithmsinan machine it's increased the timeto group the messages. Notwithstanding the reality that for little dataset time taken may be very little but for extensive datasets, the time taken is excessive. Along these lines, to evade that problem there is a need to grow such an algorithm or need to roll out the ones upgrades within the cutting-edge calculation so the hour of order lessens. Similarly, in the cutting-edge framework, there is no framework that shapes a group of messages. From the assessment, it's far likewise realized that the association of tweet based on language should be supplied. There ought to be an development of one of these framework that can arrange tweets of various dialects too.

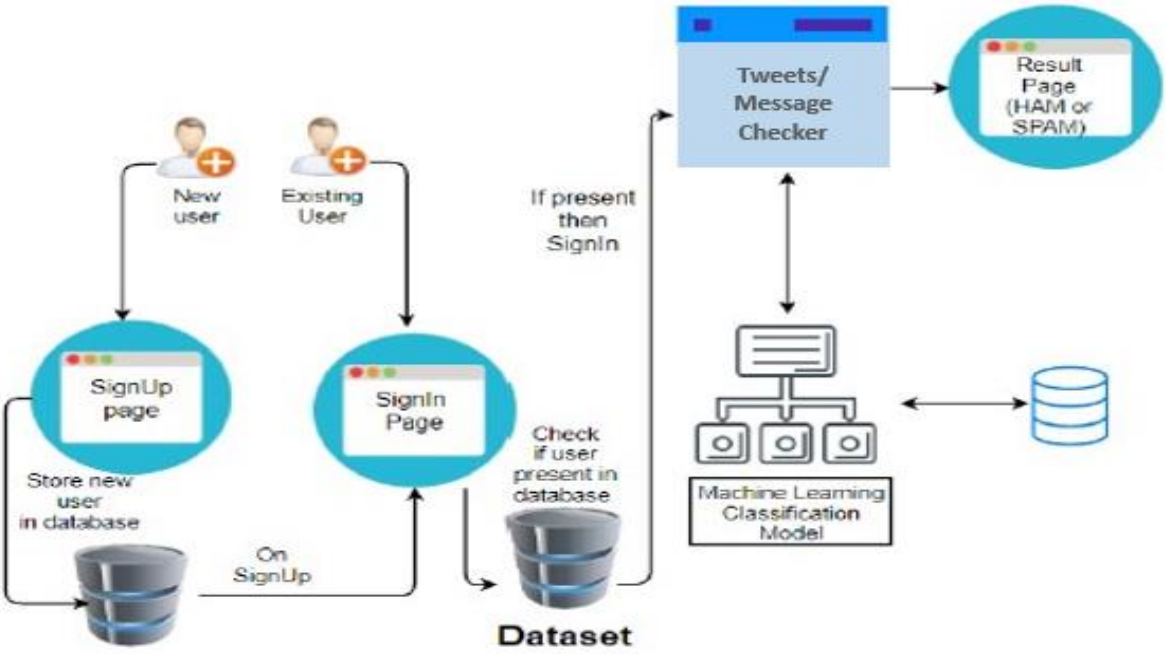
CHAPTER_3

SYSTEM ARCHITECTURE

3.1Proposed Architecture

Proposed Architecture dispays the components of client-server application.This client-server consist of SignUp and SignIn page, checker page and result page. The SignIn page is for existing users and SignUp page is for new users.After successful login,the user will be redirected to tweet/message checker page which is powered by tfid-vectoriser enanled tweet/message classification system with Logistic Regression (LR). There is sound synthesisr also placed so that people can hear for themselves of whatever the message is. It provides a better UX design. The user information is stored in the database.To use thisapplication user has to provide proper credentials.

In tweet/message checker page a textbox is given where the user will enter the tweet/message and after pressing the button, user will be redirected to result page .There is also one button to hear the text which is in the textbox.



Tweets/ Message Checker
Result Page (HAM or SPAM)
New user
Existing User
If present then SignIn
SignUp page
SignIn Page
Check if user present in database
Store new user in database
On SignUp
Dataset
Machine Learning Classification Model

Figure 3.1: Tweet/Message Classification System Architecture

The result page will display the type of tweet/message means is the entered tweet/message belongs to either Ham or Spam and also you can hear the speech of the particular text. After preprocessing of dataset, a model is being prepared with tfid-vectoriser enabled and LR algorithm in which 66.6 percent of data is used for training model and remaining 33.3 percent is used for testing. Afterpreparing model performance measures are obtained.

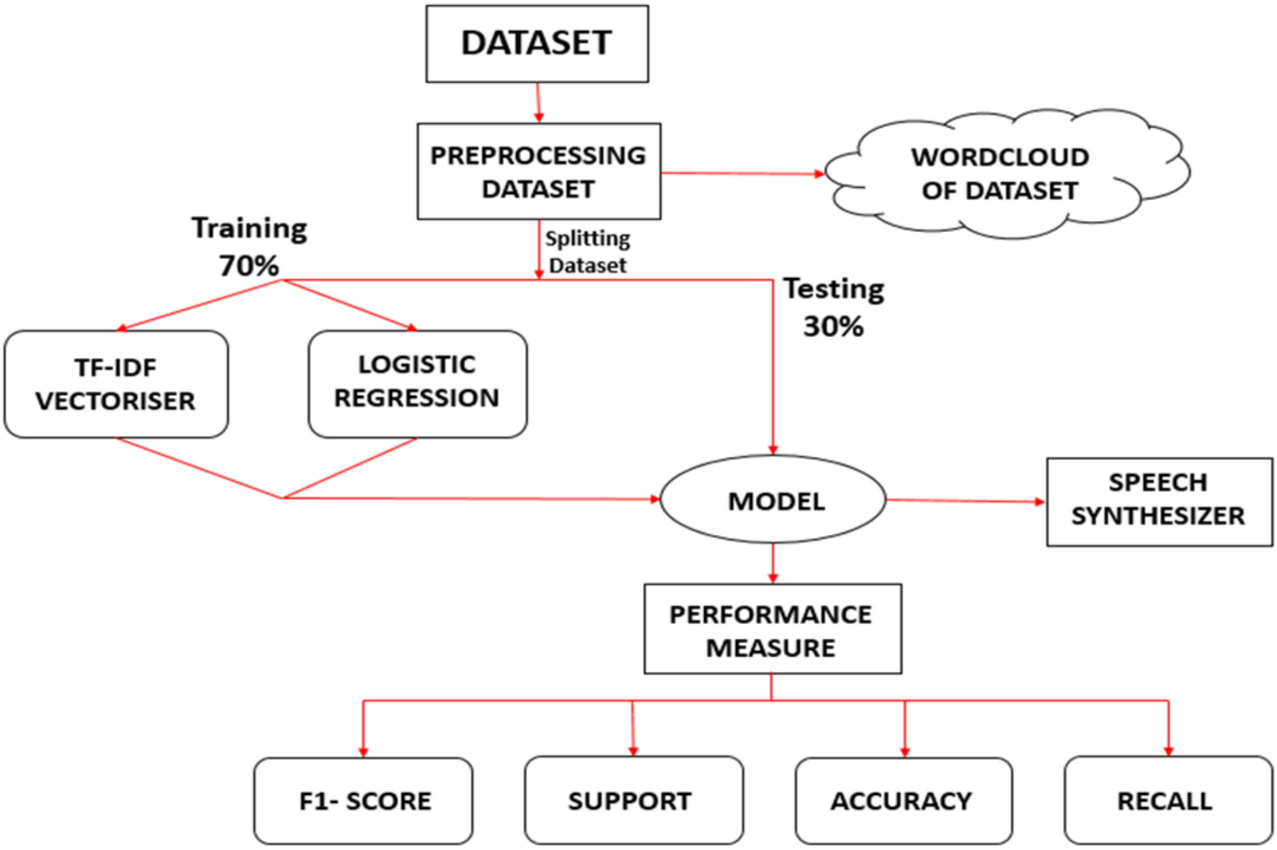


Figure 3.2: Classification Logic Architecture

CHAPTER_4

REQUIREMENT SPECIFICATION

4.1 Hardware Requirements

Any PC (laptop or desktop) will suffice as a client as long as it can compile and run python,Structured Query Language (SQL) database,and has a decent internet connection. The minimum requirements are:

- 4GB RAM
- 10GB HDD space (mainly to store the data set)
- 9 Mbps internet connection

4.2 Software Requirements

- Any Browser(Google Chrome Recommended)
- Anaconda (Python Distribution)
- SQL database-To store user information.
- Operating System-Linux,Windows or Mac
- Python 3.7 or above.

4.3 Packages Required

Below mentioned packages should be present-

- Pandas
- Sklearn
- Flask
- Sqlite3

CHAPTER_5

SYSTEM DESIGN

Here, during this area, our strateges are presented for pre-processing ofinformation, model-combination, and highlight extraction and finding execution estimates like support , F1-score, Recall value and accuracy.

5.1 Preparation of Data

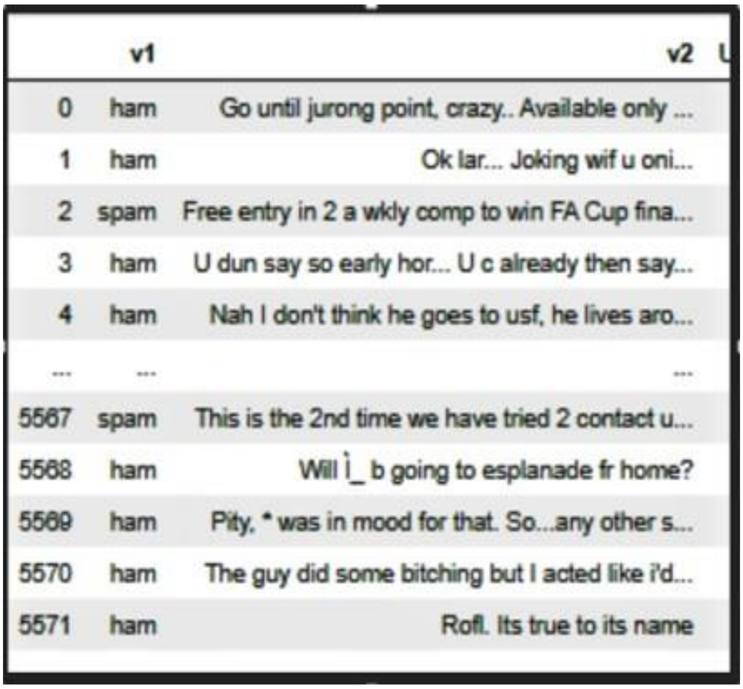
Any machine learn-ing algorithm, be it classification or regression needs a dataset to play out its importantcapacity. Along these lines, need ought to a 100 percent recovery of tweet/message from server, no matter of their area. The dataset is recovered from open-source closures to use the calculation.

When the tweet/message recovery fromthecorpus is effectively completed, it's sent to the appliance. Presently, just the difficulty of characterization is finished by filtering the content's of every tweet/message, which includes even as the type. which are parsed for extensively arranging spam and ham.The filtering procedure includes the catchphrases. The necessity is obtainable by the parameters which are perused by the calculation. The filterng and recovery structure the knowledge readiness step. The corpus contains an aggregate of about 6k messages.

The quantity of ham messages is 4.8k and therefore the quantity of spam messages of 850 within the corpus.

5.2 Analysis Of Data

The quantity ofham messages is 4.8k and also the quantity of spam messages in around 850 in the corpus.



	v1	v2
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 Ì_ 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

Figure 5.1: Number of Tweet/Message in Corpus

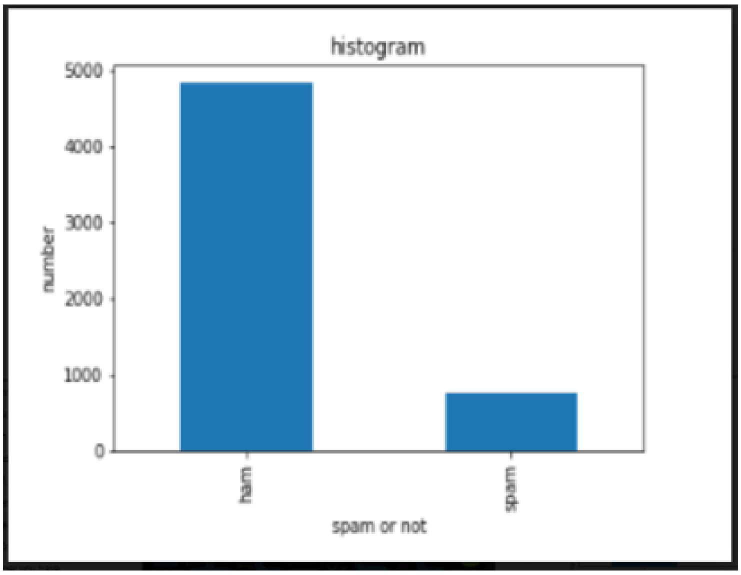


Figure 5.2: Ham and Spam Tweet/Message in Corpus

Input parameters onwhich the calculation is to be performed are thought of when the design stage is completed, After that, model go evaluation of order is employed too ’train’ dataset forcomparable examples. Any tweet/message which can be entered afterward will monitor timethat might somehow or another be utilized for extra examinations.

The exactness is straightforwardly relative to the preparation the dataset gets yet must take care with the goal that the machine isn’t over-trained. Further investgation of the result come in the wake of preparing should be reviewed. These provide a knowledge into the: ’Outliers’. Outliers demonstrate what proportion further, the outcomes go astray from the standard. The anomaly partitions the dataset into malevolent and signficant information. The right information which is acquired from the anomalies are those which will be evaluated further. It isn’t perfect that the case may be a vague-one . The calculation again emphasize to the preparation stage once the result is fixed.

5.2.1 Logistic Regression

Linear model forprediction of possibili-ties the LR method has been widely used. The expression for LR, taking h (x) stated the hypothesis, is given by:

$$h_\theta(x) = g\,(\vartheta^T x)$$

where

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

a sigmoiid-function/logistic-func. The LR model is ready by utilizing a calculation referred to as gradient ascent, in light of the likelihood of greatestprobability. In any case, LR is just too costly generally for the overwhelming and tremndous dataset and therefore the precision is undermined along these lines. stochastic gradient ascent algorithm decreases intricaicy of the iterative algorithm .It’s likewise utilized for LR model enhancement which brings about improving the iterative calculation by diminishing the intermittent puctuation. To minimize confusions, a stochastic-gradient ascent algo. is usually utilized for LRmodel improvement

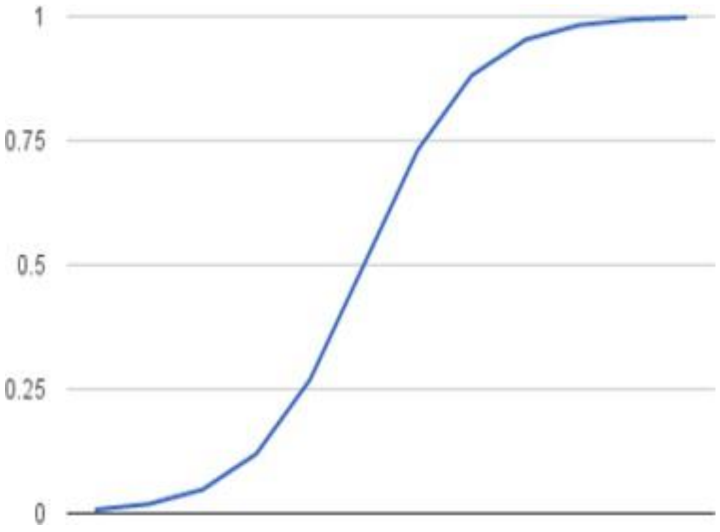
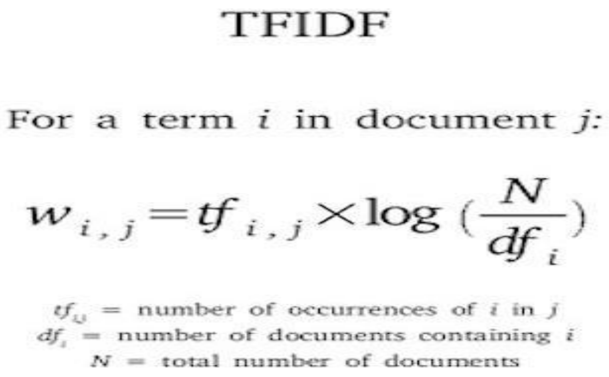


Figure 5.3: Logistic Function

5.2.2 TfidfVectorizer

The TfidfVectorizer model may be a smoothing out depiction utilized in content language takin’g care of and knowledge Retrieval (IR)). At this time, content, (for instance, a sec-tion or an expression) is addressed because the gathering (multiset) of words, disregarding discourse structure and word demand anyway keeping assortment. for information, TF-IDF is a numeric statistics that has to reflect in collection that how a word is to doc.



TFIDF

For a term i in document j:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Figure 5.4: TF-IDF Formula

TF(t) = (No. of times the term t appeared in document) / (Total no. of terms that are in document).

IDF: Inverse-Document-Frequency, which help in measuring that how important a particular term is.

We will use Tf-idf Vectoriser() as the advantage will be that it will give scores for how much times a word is appeared in sentence so it gives a float value in matrix as vector.

Before long, the TfidfVectorizer model is used as a gadget for feature . Resulting to change substance to "pack of words", we will learn various measures to depict the substance. The foremost broadly perceived kind of traits or features decided from the Bag-of-words is repeat, to be explicit, the events a term appears within the substance.

5.3 Final Assessment

At long last, before the yield is appeared, it should be surveyed once and for all. At that time this is often ordered or separated. The dataset has been prepared on the calculation we require. The continual outcome are often gotten by recovering the client’s messages. The precision are often advanced by contrasting the result and therefore the ongoing outcomes. The parameters are additionally held for the calculation. Further, this may empower us in investigating the vulnerability of the acquired result and improve the degree of exactness for shifting and prioritization.

CHAPTER_6

MODULES

6.1 Module 1-Sign Up

This module is used for adding new users to use the Tweets/message classification web application. It is connected with a SQL database. It stores email, name and password for new users. This module renders the signup.html file which helps user to easily access this web application and provides a visually attractive experience to the users. After successful signUp, the user is navigated to signin.html page.

6.2 Module 2-Sign In

This module is used for authorization and authentication. Only users which are added in database can signIn. This module renders signin.html page which contains email and password fields .This module ensures that the entered credentials are present in database. If the credentials are in database and proper, then user will be redirected to checker.html page otherwise alerts will be shown to user based on the error that “Something went Wrong”.

6.3 Module 3-Checker

This module renders checker.html file. checker.html file consist of a textbox area where theuser enters the tweet/message. This page also consist of predict button. On pressing, "PREDICT" button the entered tweet/message is sent to "predict" page using POST request. Also we add a Speech button. On pressing, “SPEECH” button the entered tweet/message is sent to “speech” function where the particular message converted to speech from text and the sound will be heared on the same page only.

6.4 Module 4-Predict

This module contains the main logic behind this tweet/message classification system. This module contains the LR algorithm powered by TfidfVectoriser. The dataset used is "spam.csv"to tran the model. 66.6 percent data of dataset i.e. 2/3rd is used for training the model and remaning 33.3 percent i.e. 1/3rd is used for testing. The data obtained from module 3-Checker is passed to this model and based on type of tweet/message entered in textbox either "Ham" or "Spam" is being displayed on webpage by the result.html page which is rendered by this module.

CHAPTER_7

BENEFITS

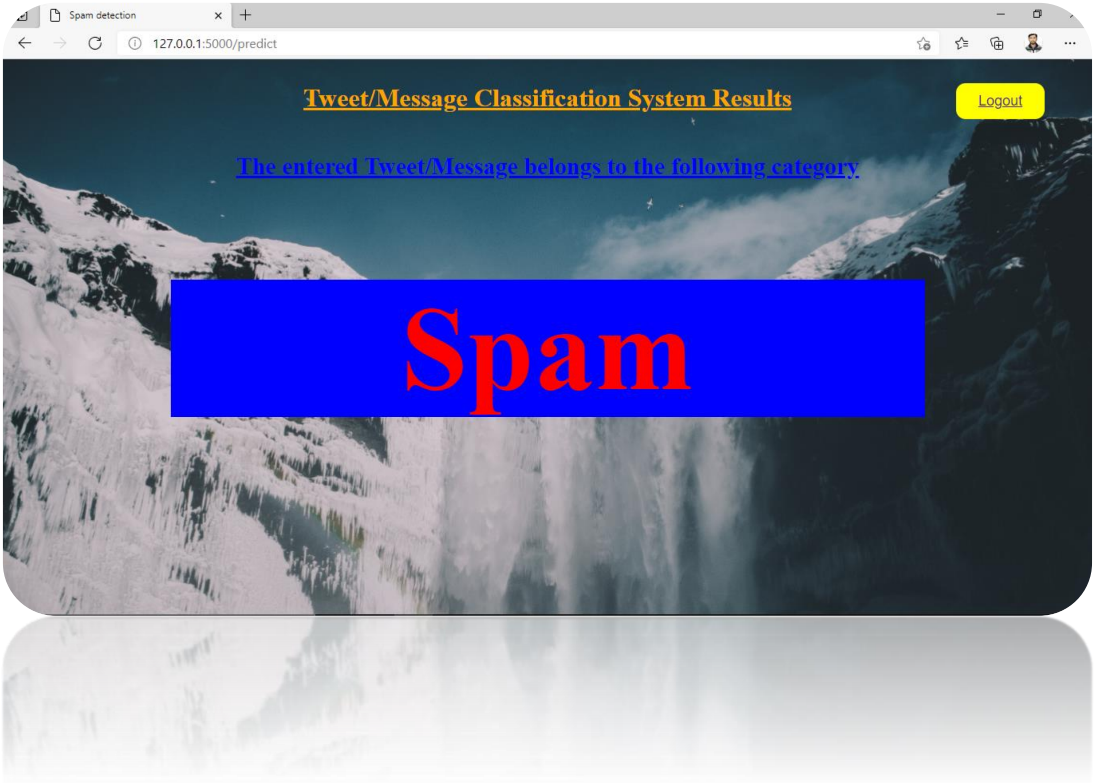
1. Accuracy - The classification of tweet/message done by the system is highly accurate andprecise as the accuracy of the trained model is 97 percent.
2. Easy to Use - The entire web application is easy to use because of the graphical interface made using HTML and Cascading Style Sheet (CSS) files.Providing input is very easy and visually attractive as well as obtained output is very easy to understand.
3. Secure System - The entire system is secure to use.The system has signUp and SignIn page.Hence no unauthorised user can use this application.This provides safety to users and also beneficial for the system to store data.
4. User Friendly - More Comfortable to use when speech is added to the project and that makes it much easier for the user to operate.
5. UX Design - Good UX Design helps user to use the tool in a better way as it provides a customer-centric experience.

CHAPTER_8

RESULTS

After entering tweet/message in textbox and pressing "Predict" button.

Input 1-, all 4 FREE! bx420-ip4-5we. 150pm. Dont missout! Congrats! 1 yearspecial cinemapass for 2 is been yours. call 09061209465 now! C SuprmanV, Matrix-3, StarWars-3,



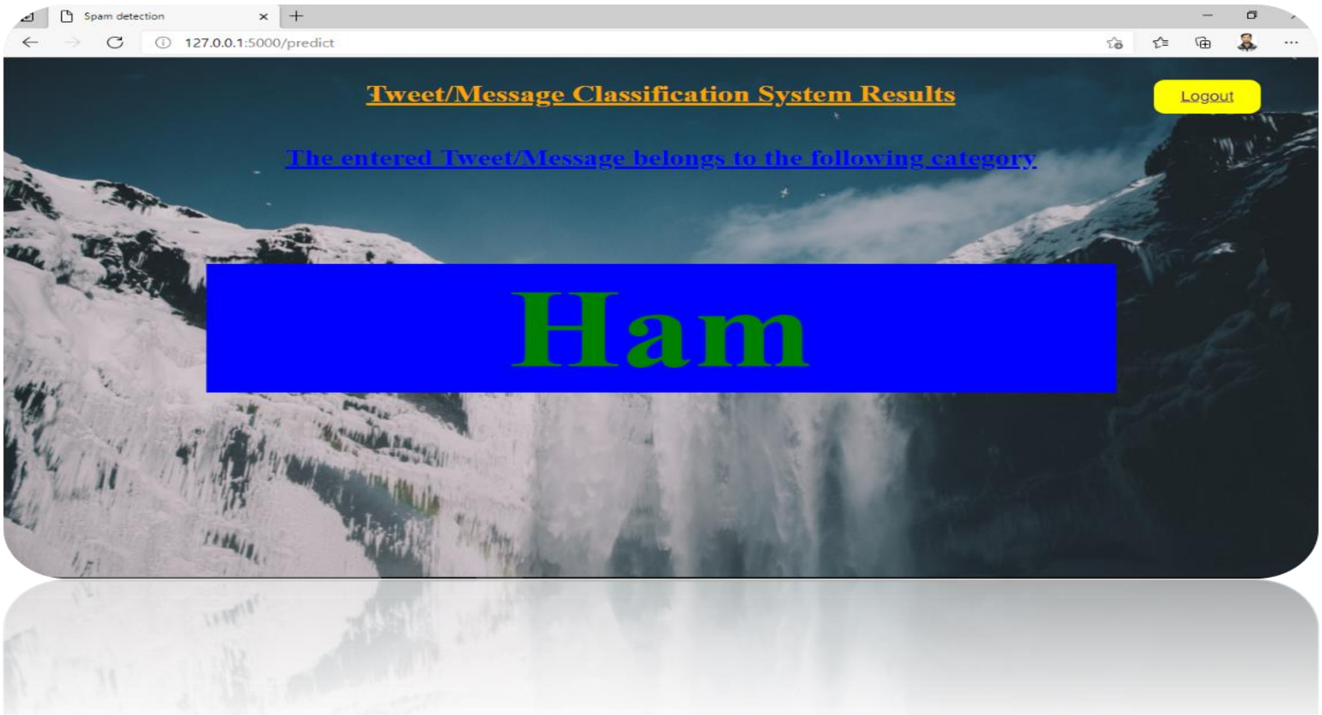
Tweet/Message Classification System Results

The entered Tweet/Message belongs to the following category

Spam

Figure 8.1: Spam tweets/message Test Result

Input 2- hey how are u, I m fine TY



Tweet/Message Classification System Results

The entered Tweet/Message belongs to the following category

Ham

Figure 8.2: Ham Tweet/message Test Result

WordCloud Obtained- The following wordcloud is got from the corpus. “are a part of the organization’s daily work which is issued by tool by the social security authorities to see if any activities like ,frauds,terror.It displays size of words per thefrequency of a word appearing in data.It is easy way to analyze by just looking.It is visu-ally attractive and easy to understand.It can be also used to perform sentiment analysis.

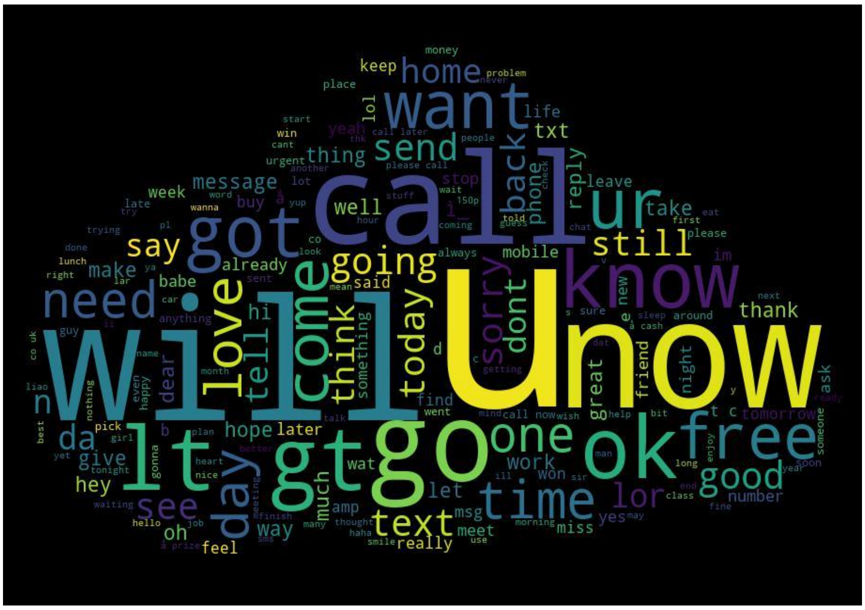


Figure 8.3: WordCloud Obtained

CHAPTER_9

CONCLUSION

After entering tweet/message, the results obtained are correct i.e. whether Ham or Spam..A per-formance measures are as follows. In this model using TfidfVectorizer along with LR,

we found the accuracy of 97 percent for classify tweet/message into ham or spam. In this model, F1-score for ham- 0.98 and spam has- 0.86. Recall-value for ham- 0.96 and spam has- 0.99

Table 9.1: Performance Measures

Type	Precision	Recall	F1-Score	Support
Ham	1.00	.96	.98	1648
Spam	.77	.99	.86	191

CHAPTER_10

FUTURE ENHANCEMENTS

1. A history storing mechanism can be deployed in this, so that email entered by the user can be saved as a history with it results predicted by the LR model and if the same tweet/message is again entered then the same results can be shown.This will improve its performanceand also save time to predict the results.

2. 2.User Interface (UI) can be improved because currently the UI is made up of HTML and CSS.The UI can be improved by using more robust and secure JavaScript (JS) library like React.js, Angular.js, Vue.js etc.

3. More security checks can be added like One-Time Password (OTP) or Audio or Picture CAPTCHA as to ensure it cannot be used for illegal activities.

4. Speech-to-text can also be added so that the user has to just speak and the message get transcript to text and got paste to textbox which can be checked as well. This will help in model to becoming more userfriendly.

REFERENCES

[1] HadiOtrok ,BahiaHalawi,AzzamMourad ,ErnestoDamiani ; An Ontology-Based TweetSpam Approach Detection ; Year-2018

[2] AbdullahTalah, Resulkara ; A Spam-Detection Methodson Twitter ; Year-2017

[3] Mohd.SaalimJamal, HimankGupta & MaunendraSankarDesarkar , Sreekanth Madisetty; Framework of Real-Time SpamDetect inTwitter ; Year-2018

[4] S.ShajunNisha, M.MohamadSathik, M.Gayathri ;TwitterSentiment Analysis ; Year-2020

[5] HadeelAl-Negheimish, Nora Al-Twairesh; Deep andsurface Features Ensemble for SentimentAnalysis of ArabicTweets ; Year-2019

[6] Yoo, S Ph.D. Carnegie Mellon University ; MachineLearning methods for personalizedemail prioritization: year-2010

[7] Sasikumaran SreedharanAbeer Alsadoon,P.W.C. Prasad,M. K. Chae;Spam filtering emailclassifier (SFECM) using gainandgraph miningalgorithm;Year(2017)

[8] B.NiranjanaKrupa, M.S.Dhananjaya;R. Sushma;Kannada speechtotext conversion: A novel approach,IEEE-2016

[9] Sk Golam Saroar,Md Mosfaiul,Alam SebastianRomy Gomes Telot,Behroz Newaz Khan,AmitabhaChakrabarty ; A comparativeapproach to email classification usinghidden Markov model andNaive Bayes classifier ; Year-2017

[10] P Punde,R Wagh;Survey on sentimentanalysis using twitterdataset : Year(2018)

[11] P Rosso,E Fersini,M.Anzovino;Automatic identification and classificationof misogynistic languageon twitter; Year(2018)

[12] AlexHaiWang;machine learning forthe Detection of SpaminTwitter Networks, springer; Year(2012)

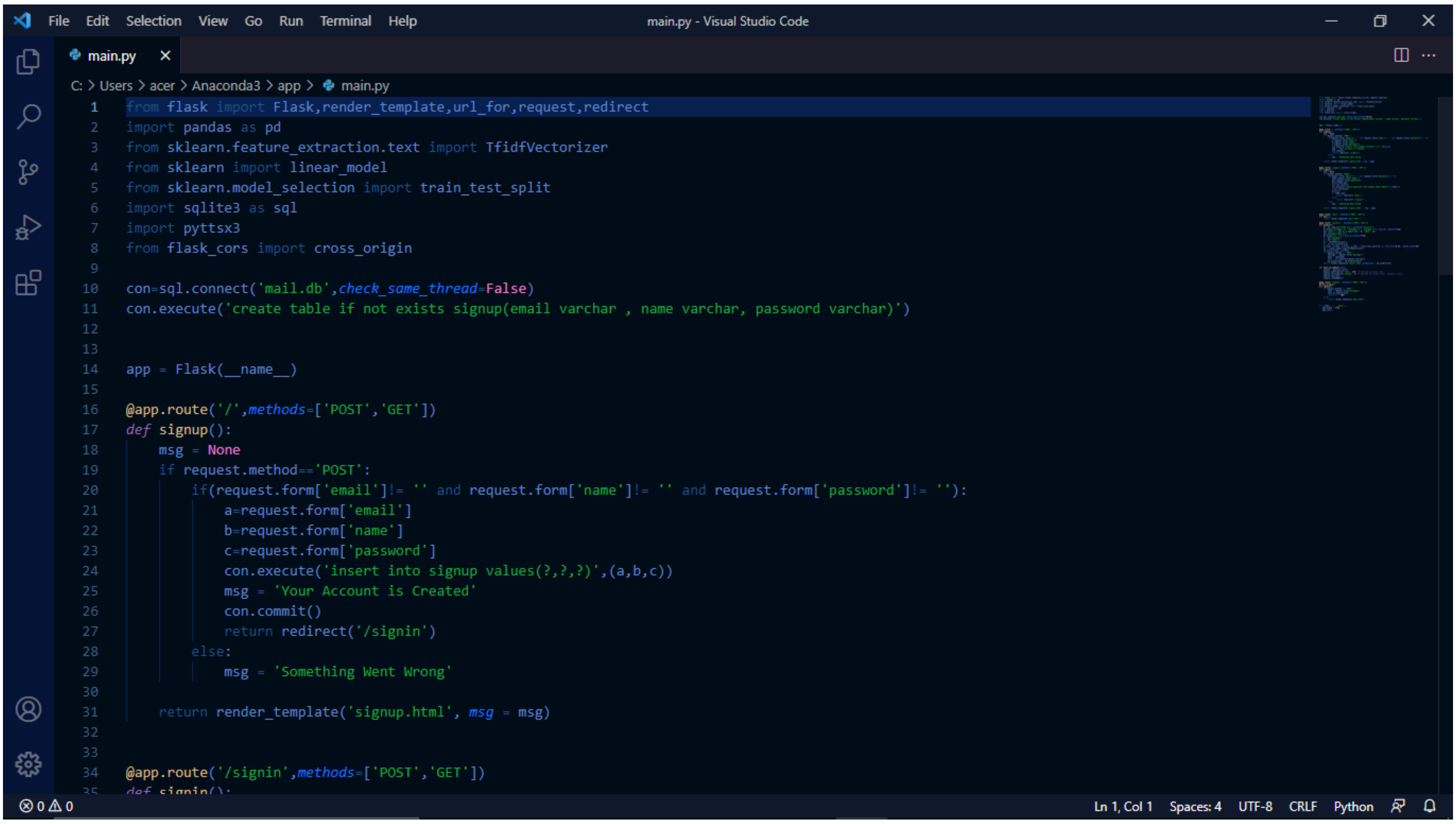
[13] S Nepal,R Nugroho,J Yang,C Paris; A survey ofrecent methodson deriving topicsfrom Twitter: algorithm toevaluation; Year(2020)

[14] V Kharde, P Sonawane ;Sentimentanalysis oftwitter data: asurvey of techniques-2016

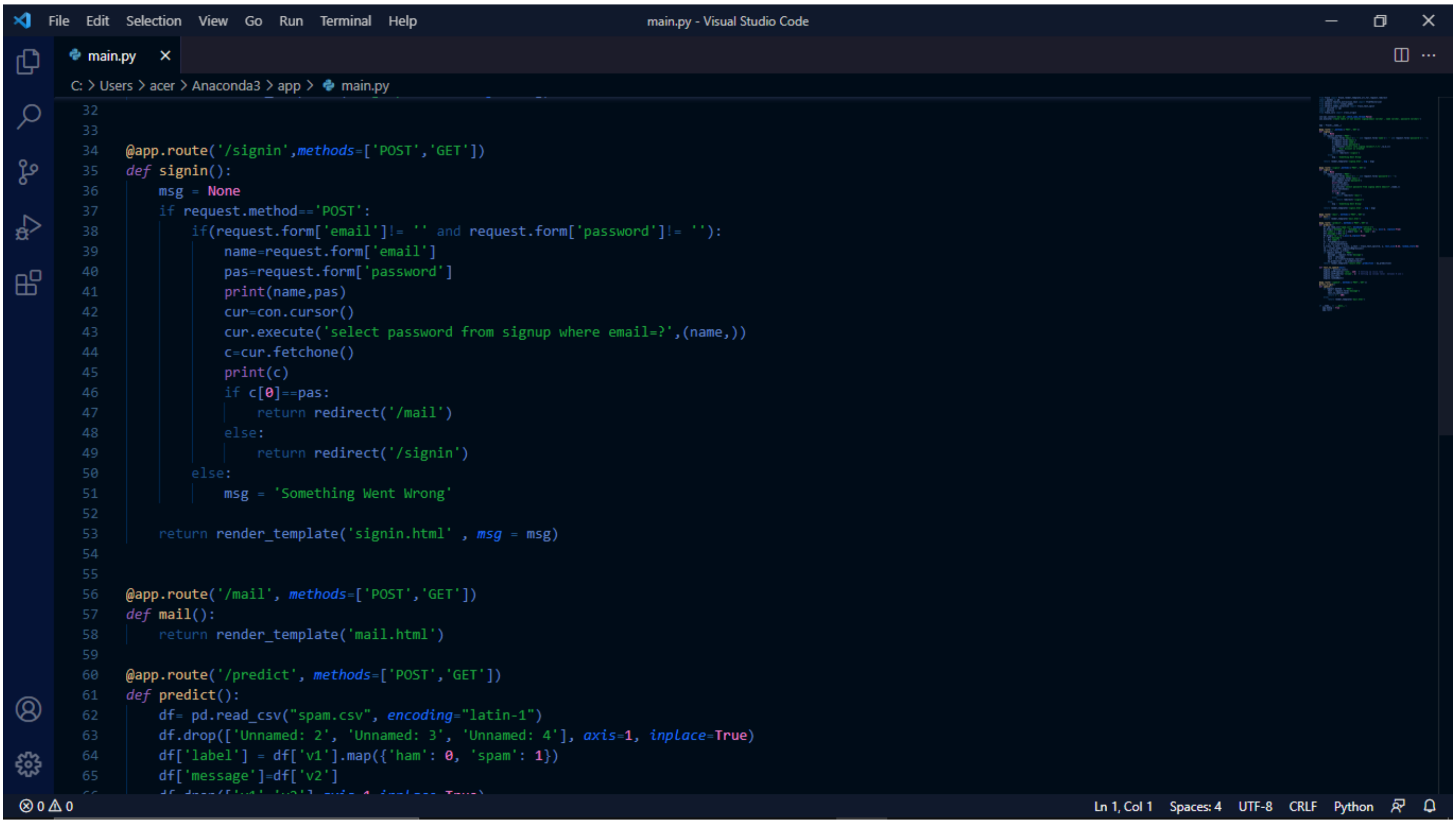
CHAPTER_11

APPENDICES

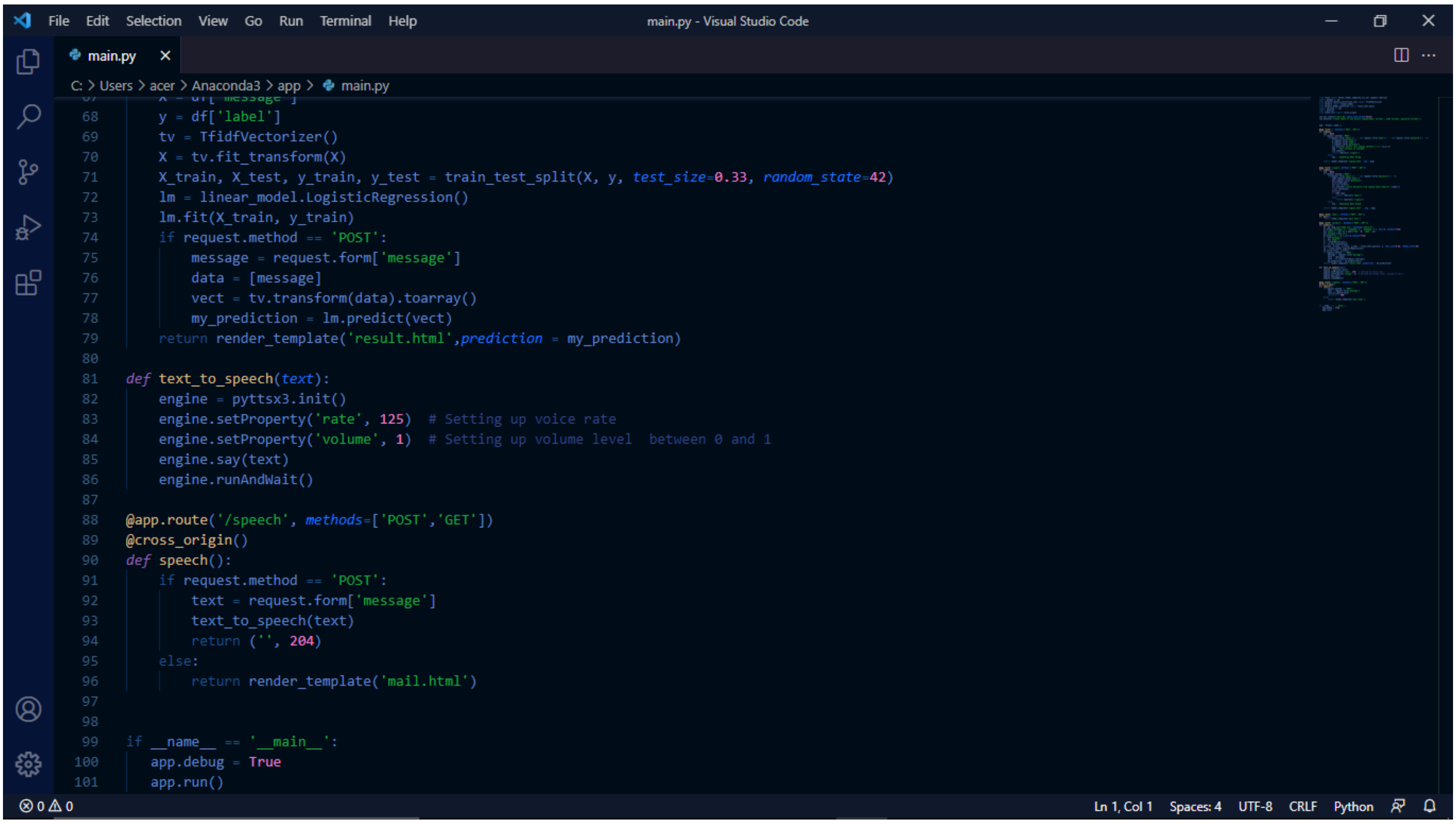
main.py



```python
from flask import Flask,render_template,url_for,request,redirect
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn import linear_model
from sklearn.model_selection import train_test_split
import sqlite3 as sql
import pyttsx3
from flask_cors import cross_origin

con=sql.connect('mail.db',check_same_thread=False)
con.execute('create table if not exists signup(email varchar , name varchar, password varchar)')


app = Flask(__name__)

@app.route('/',methods=['POST','GET'])
def signup():
    msg = None
    if request.method=='POST':
        if(request.form['email']!= '' and request.form['name']!= '' and request.form['password']!= ''):
            a=request.form['email']
            b=request.form['name']
            c=request.form['password']
            con.execute('insert into signup values(?,?,?)',(a,b,c))
            msg = 'Your Account is Created'
            con.commit()
            return redirect('/signin')
        else:
            msg = 'Something Went Wrong'

    return render_template('signup.html', msg = msg)


@app.route('/signin',methods=['POST','GET'])
def signin():
```

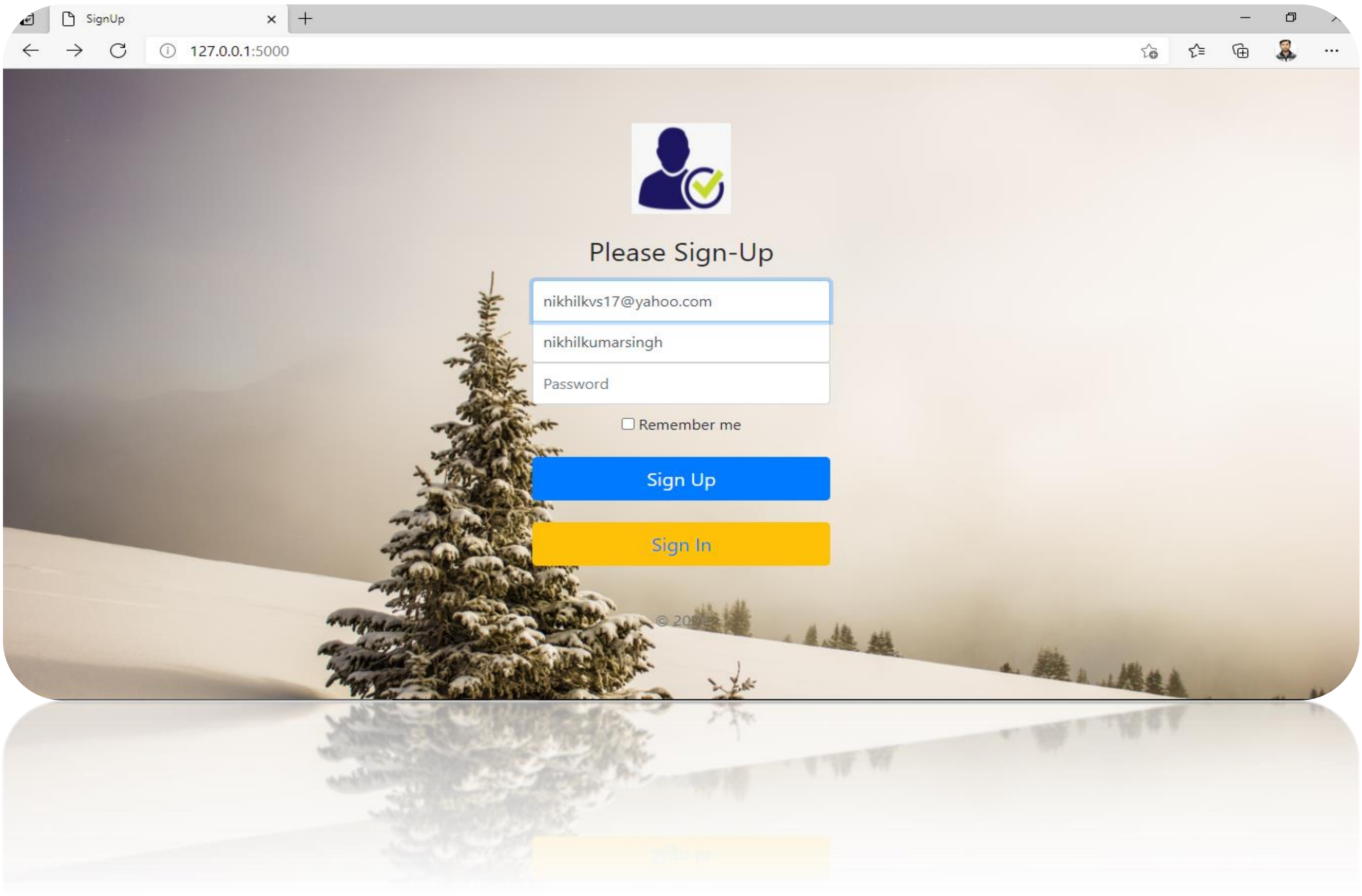


```python

@app.route('/signin',methods=['POST','GET'])
def signin():
    msg = None
    if request.method=='POST':
        if(request.form['email']!= '' and request.form['password']!= ''):
            name=request.form['email']
            pas=request.form['password']
            print(name,pas)
            cur=con.cursor()
            cur.execute('select password from signup where email=?',(name,))
            c=cur.fetchone()
            print(c)
            if c[0]==pas:
                return redirect('/mail')
            else:
                return redirect('/signin')
        else:
            msg = 'Something Went Wrong'

    return render_template('signin.html' , msg = msg)


@app.route('/mail', methods=['POST','GET'])
def mail():
    return render_template('mail.html')

@app.route('/predict', methods=['POST','GET'])
def predict():
    df= pd.read_csv("spam.csv", encoding="latin-1")
    df.drop(['Unnamed: 2', 'Unnamed: 3', 'Unnamed: 4'], axis=1, inplace=True)
    df['label'] = df['v1'].map({'ham': 0, 'spam': 1})
    df['message']=df['v2']
    df.drop(['v1','v2'],axis=1,inplace=True)
```



```python
    X = df['message']
    y = df['label']
    tv = TfidfVectorizer()
    X = tv.fit_transform(X)
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)
    lm = linear_model.LogisticRegression()
    lm.fit(X_train, y_train)
    if request.method == 'POST':
        message = request.form['message']
        data = [message]
        vect = tv.transform(data).toarray()
        my_prediction = lm.predict(vect)
    return render_template('result.html',prediction = my_prediction)

def text_to_speech(text):
    engine = pyttsx3.init()
    engine.setProperty('rate', 125)  # Setting up voice rate
    engine.setProperty('volume', 1)  # Setting up volume level  between 0 and 1
    engine.say(text)
    engine.runAndWait()

@app.route('/speech', methods=['POST','GET'])
@cross_origin()
def speech():
    if request.method == 'POST':
        text = request.form['message']
        text_to_speech(text)
        return ('', 204)
    else:
        return render_template('mail.html')


if __name__ == '__main__':
   app.debug = True
   app.run()
```

signup.html



Please Sign-Up

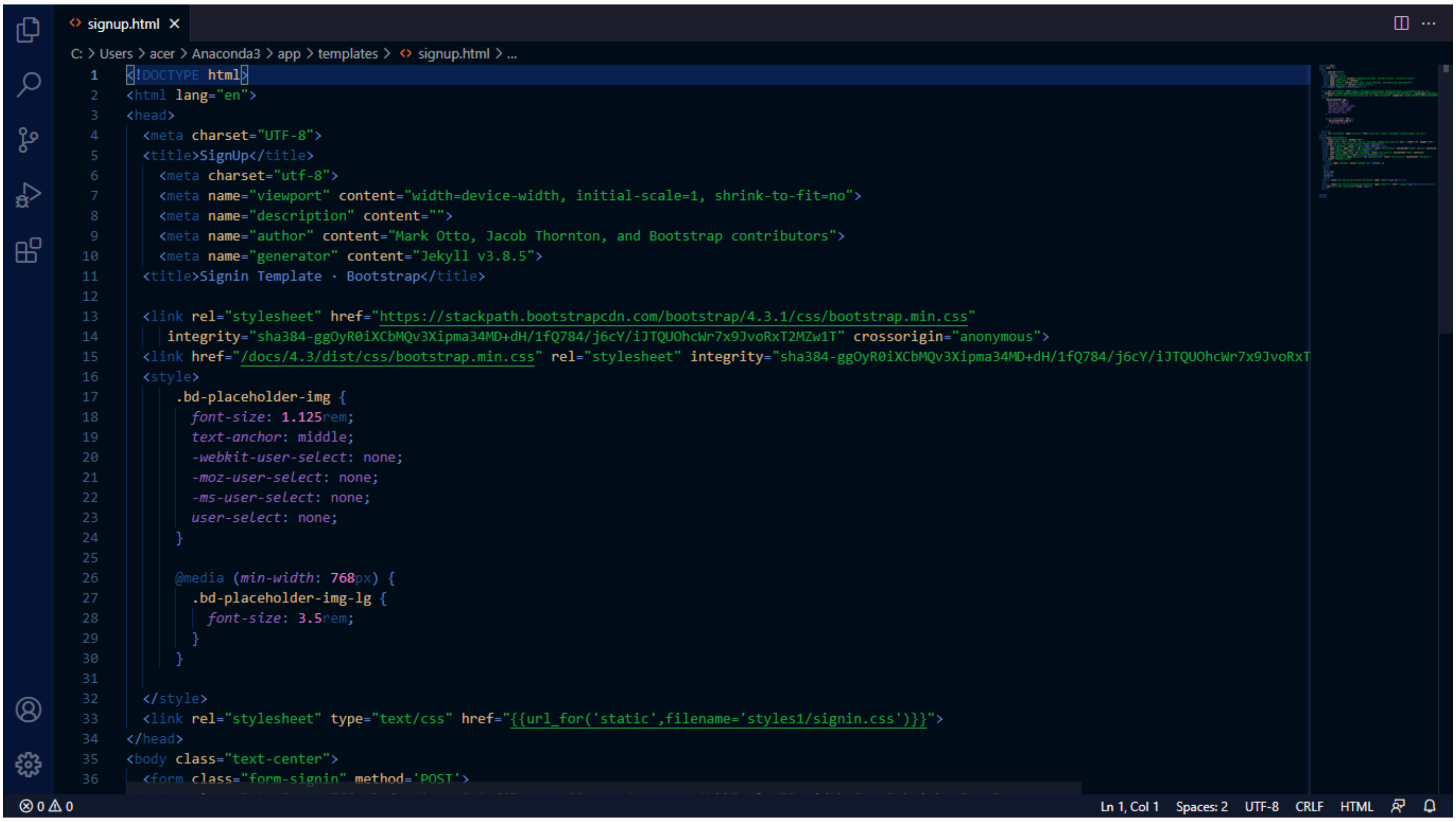
nikhilkvs17@yahoo.com

nikhilkumarsingh

Password

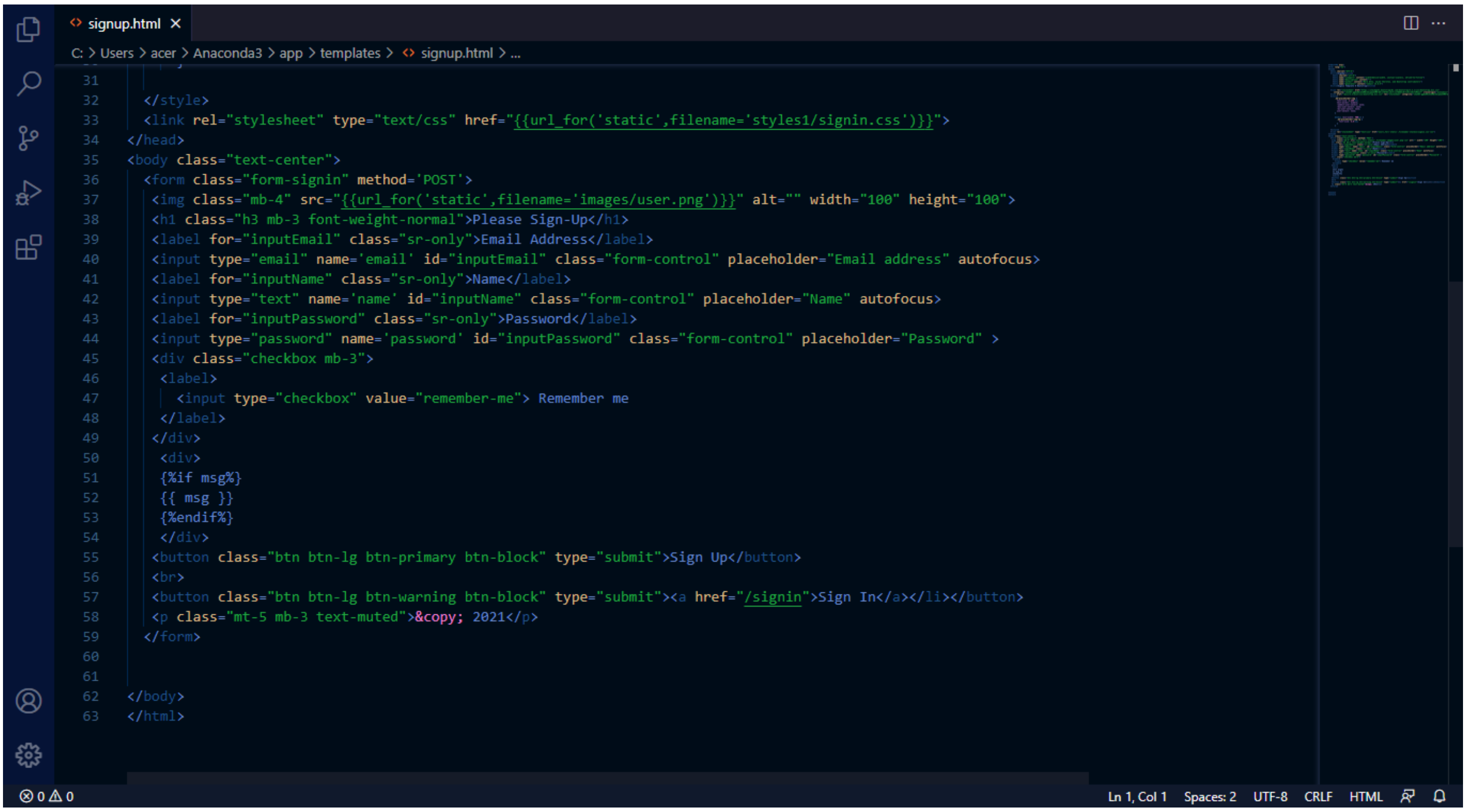
□ Remember me

Figure 11.1: Signup page



signup.html
C: > Users > acer > Anaconda3 > app > templates > signup.html > ...
⊗ 0 ⚠ 0
Ln 1, Col 1
Spaces: 2
UTF-8
CRLF
HTML

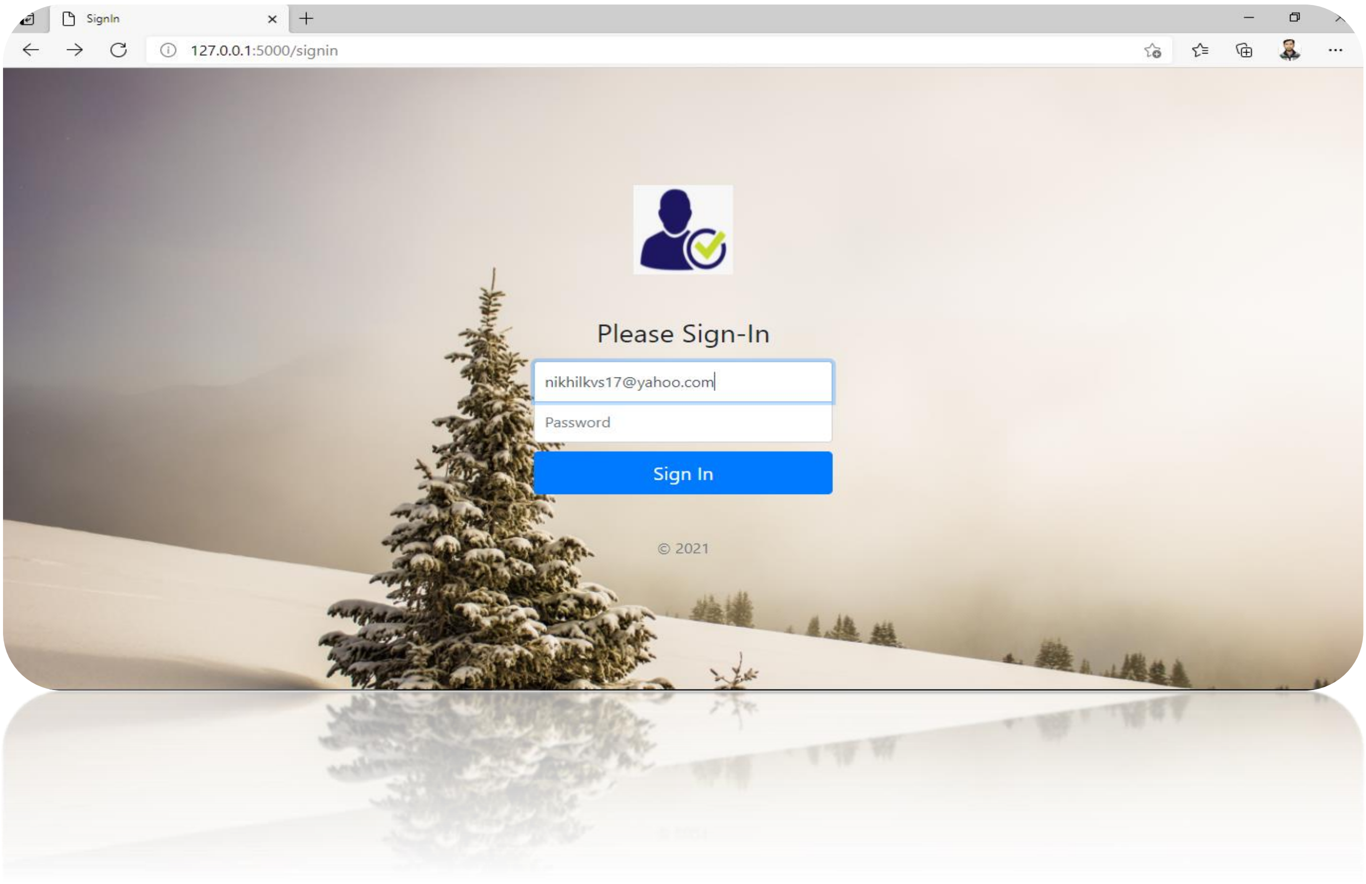
```
<!DOCTYPE html>
<html lang="en">
<head>
  <meta charset="UTF-8">
  <title>SignUp</title>
    <meta charset="utf-8">
    <meta name="viewport" content="width=device-width, initial-scale=1, shrink-to-fit=no">
    <meta name="description" content="">
    <meta name="author" content="Mark Otto, Jacob Thornton, and Bootstrap contributors">
    <meta name="generator" content="Jekyll v3.8.5">
  <title>Signin Template · Bootstrap</title>

  <link rel="stylesheet" href="https://stackpath.bootstrapcdn.com/bootstrap/4.3.1/css/bootstrap.min.css"
     integrity="sha384-ggOyR0iXCbMQv3Xipma34MD+dH/1fQ784/j6cY/iJTQUOhcWr7x9JvoRxT2MZw1T" crossorigin="anonymous">
  <link href="/docs/4.3/dist/css/bootstrap.min.css" rel="stylesheet" integrity="sha384-ggOyR0iXCbMQv3Xipma34MD+dH/1fQ784/j6cY/iJTQUOhcWr7x9JvoRxT
  <style>
      .bd-placeholder-img {
        font-size: 1.125rem;
        text-anchor: middle;
        -webkit-user-select: none;
        -moz-user-select: none;
        -ms-user-select: none;
        user-select: none;
      }

      @media (min-width: 768px) {
        .bd-placeholder-img-lg {
          font-size: 3.5rem;
        }
      }

  </style>
  <link rel="stylesheet" type="text/css" href="{{url_for('static',filename='styles1/signin.css')}}">
</head>
<body class="text-center">
  <form class="form-signin" method='POST'>
```



```
   </style>
   <link rel="stylesheet" type="text/css" href="{{url_for('static',filename='styles1/signin.css')}}">
 </head>
 <body class="text-center">
   <form class="form-signin" method='POST'>
    <img class="mb-4" src="{{url_for('static',filename='images/user.png')}}" alt="" width="100" height="100">
    <h1 class="h3 mb-3 font-weight-normal">Please Sign-Up</h1>
    <label for="inputEmail" class="sr-only">Email Address</label>
    <input type="email" name='email' id="inputEmail" class="form-control" placeholder="Email address" autofocus>
    <label for="inputName" class="sr-only">Name</label>
    <input type="text" name='name' id="inputName" class="form-control" placeholder="Name" autofocus>
    <label for="inputPassword" class="sr-only">Password</label>
    <input type="password" name='password' id="inputPassword" class="form-control" placeholder="Password" >
    <div class="checkbox mb-3">
     <label>
       <input type="checkbox" value="remember-me"> Remember me
     </label>
    </div>
     <div>
     {%if msg%}
     {{ msg }}
     {%endif%}
     </div>
    <button class="btn btn-lg btn-primary btn-block" type="submit">Sign Up</button>
    <br>
    <button class="btn btn-lg btn-warning btn-block" type="submit"><a href="/signin">Sign In</a></li></button>
    <p class="mt-5 mb-3 text-muted">&copy; 2021</p>
   </form>


 </body>
 </html>
```

signup.html
C: > Users > acer > Anaconda3 > app > templates > signup.html > ...
⊗ 0 ⚠ 0
Ln 1, Col 1 Spaces: 2 UTF-8 CRLF HTML

signin.html



127.0.0.1:5000/signin

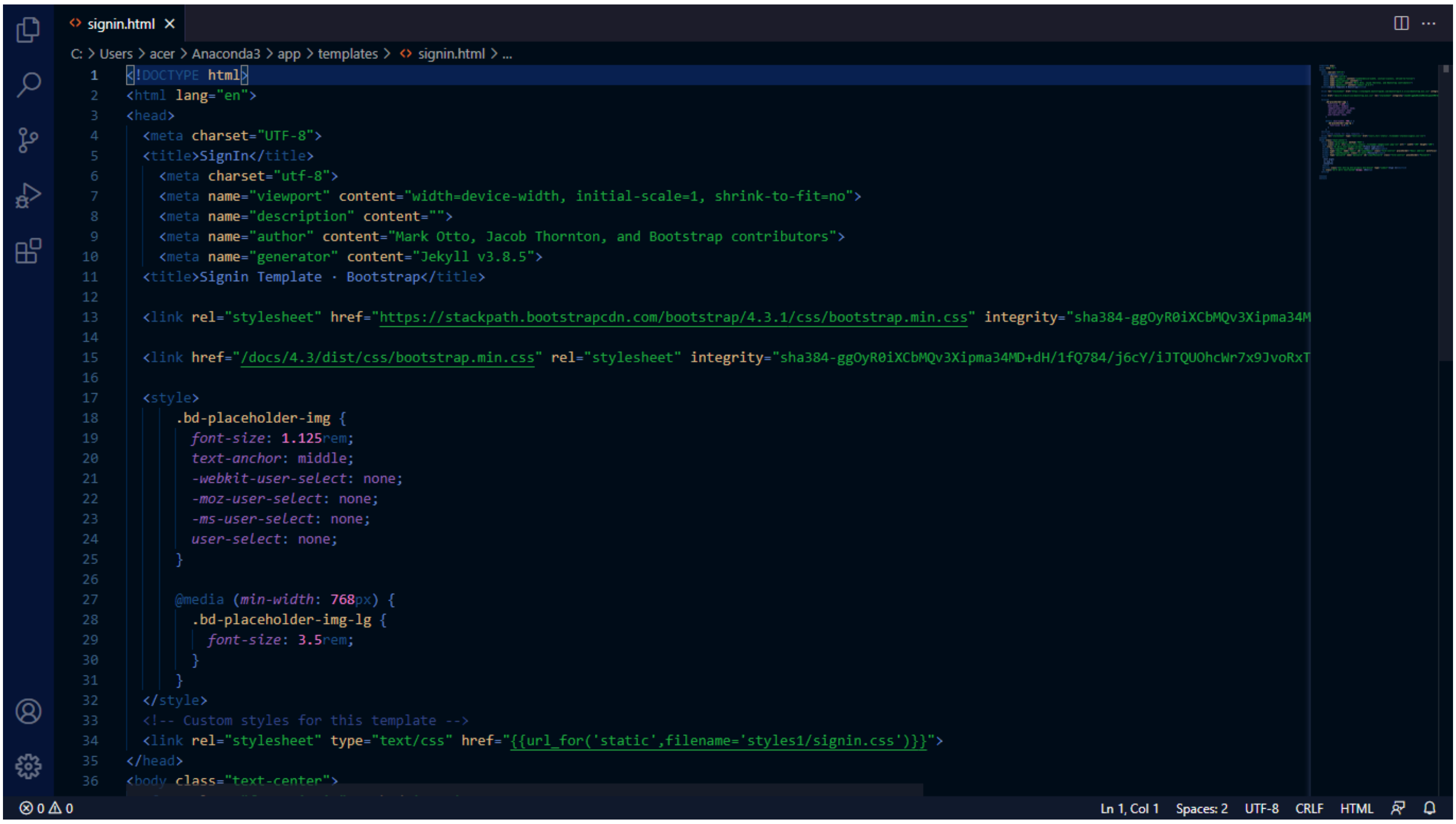
Please Sign-In

nikhilkvs17@yahoo.com

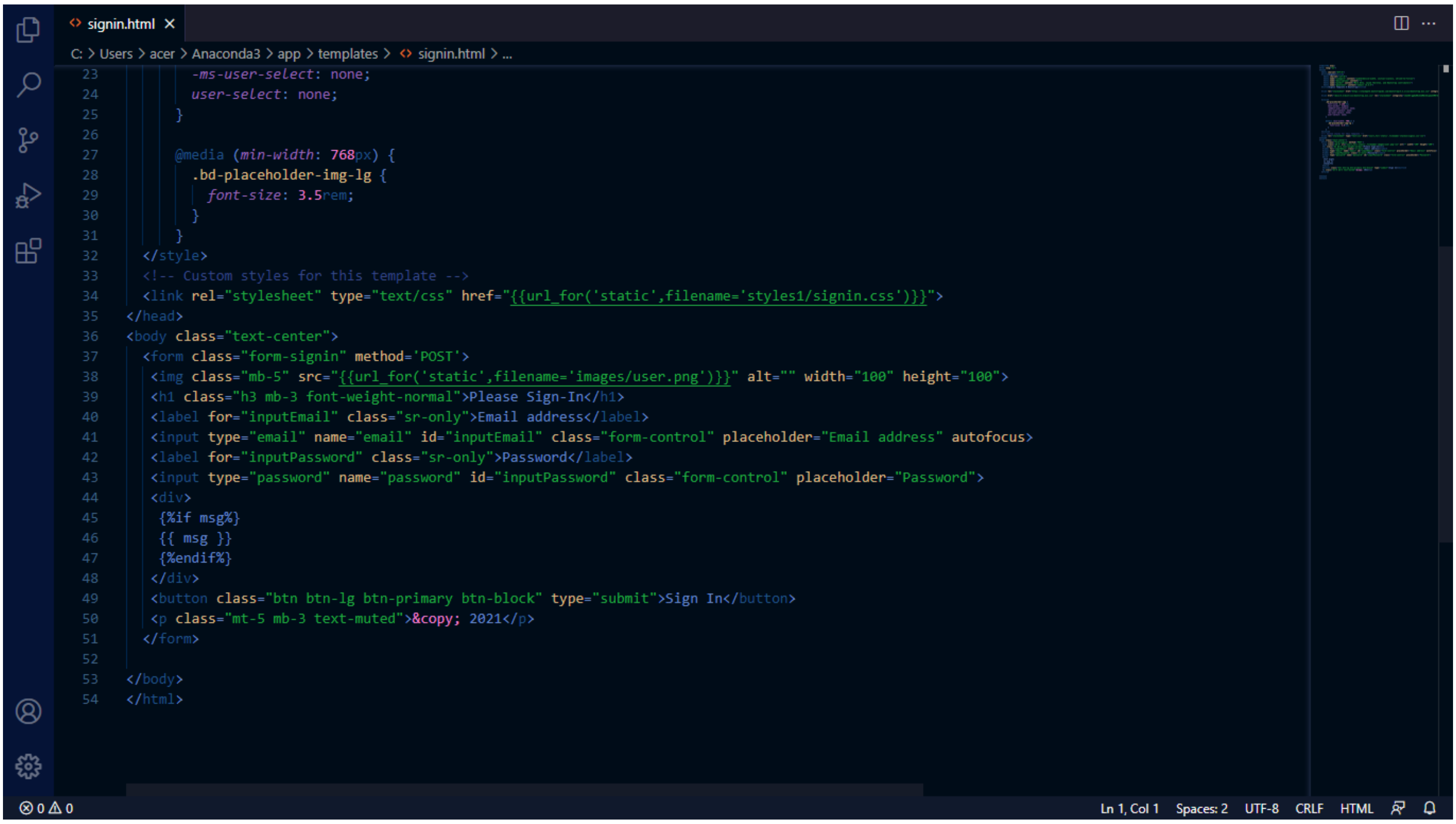
Password

© 2021

Figure 11.2: Signin page


signin.html
C: > Users > acer > Anaconda3 > app > templates > signin.html > ...
⊗ 0 ⚠ 0
Ln 1, Col 1
Spaces: 2
UTF-8
CRLF
HTML

```
<!DOCTYPE html>
<html lang="en">
<head>
  <meta charset="UTF-8">
  <title>SignIn</title>
    <meta charset="utf-8">
    <meta name="viewport" content="width=device-width, initial-scale=1, shrink-to-fit=no">
    <meta name="description" content="">
    <meta name="author" content="Mark Otto, Jacob Thornton, and Bootstrap contributors">
    <meta name="generator" content="Jekyll v3.8.5">
  <title>Signin Template · Bootstrap</title>

  <link rel="stylesheet" href="https://stackpath.bootstrapcdn.com/bootstrap/4.3.1/css/bootstrap.min.css" integrity="sha384-ggOyR0iXCbMQv3Xipma34M

  <link href="/docs/4.3/dist/css/bootstrap.min.css" rel="stylesheet" integrity="sha384-ggOyR0iXCbMQv3Xipma34MD+dH/1fQ784/j6cY/iJTQUOhcWr7x9JvoRxT

  <style>
      .bd-placeholder-img {
        font-size: 1.125rem;
        text-anchor: middle;
        -webkit-user-select: none;
        -moz-user-select: none;
        -ms-user-select: none;
        user-select: none;
      }

      @media (min-width: 768px) {
        .bd-placeholder-img-lg {
          font-size: 3.5rem;
        }
      }
  </style>
  <!-- Custom styles for this template -->
  <link rel="stylesheet" type="text/css" href="{{url_for('static',filename='styles1/signin.css')}}">
</head>
<body class="text-center">
```


signin.html
C: > Users > acer > Anaconda3 > app > templates > signin.html > ...
⊗ 0 ⚠ 0
Ln 1, Col 1
Spaces: 2
UTF-8
CRLF
HTML

```
        -ms-user-select: none;
        user-select: none;
      }

      @media (min-width: 768px) {
        .bd-placeholder-img-lg {
          font-size: 3.5rem;
        }
      }
  </style>
  <!-- Custom styles for this template -->
  <link rel="stylesheet" type="text/css" href="{{url_for('static',filename='styles1/signin.css')}}">
</head>
<body class="text-center">
  <form class="form-signin" method='POST'>
   <img class="mb-5" src="{{url_for('static',filename='images/user.png')}}" alt="" width="100" height="100">
   <h1 class="h3 mb-3 font-weight-normal">Please Sign-In</h1>
   <label for="inputEmail" class="sr-only">Email address</label>
   <input type="email" name="email" id="inputEmail" class="form-control" placeholder="Email address" autofocus>
   <label for="inputPassword" class="sr-only">Password</label>
   <input type="password" name="password" id="inputPassword" class="form-control" placeholder="Password">
   <div>
    {%if msg%}
    {{ msg }}
    {%endif%}
   </div>
   <button class="btn btn-lg btn-primary btn-block" type="submit">Sign In</button>
   <p class="mt-5 mb-3 text-muted">&copy; 2021</p>
  </form>

</body>
</html>
```

checker.html

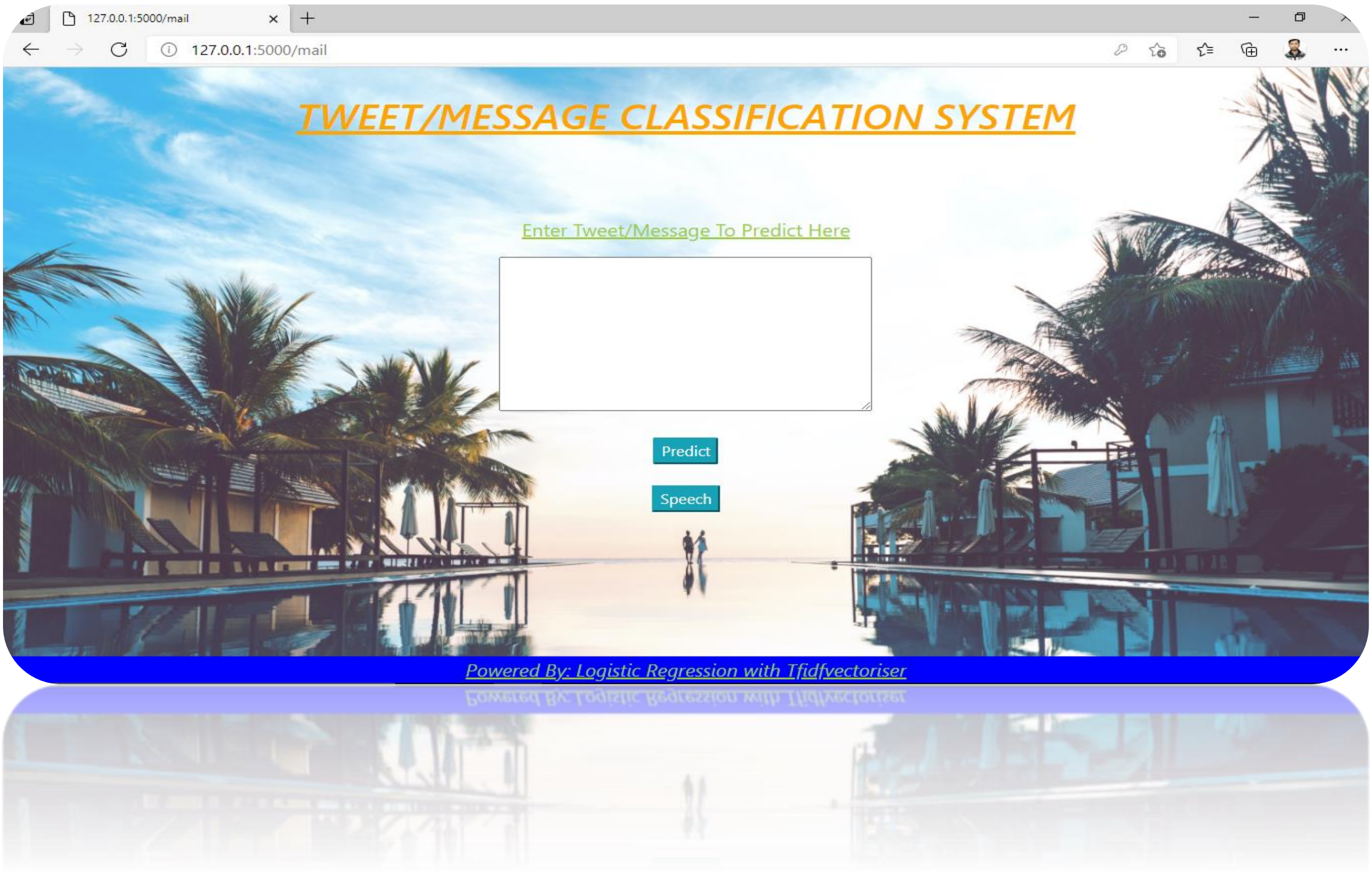
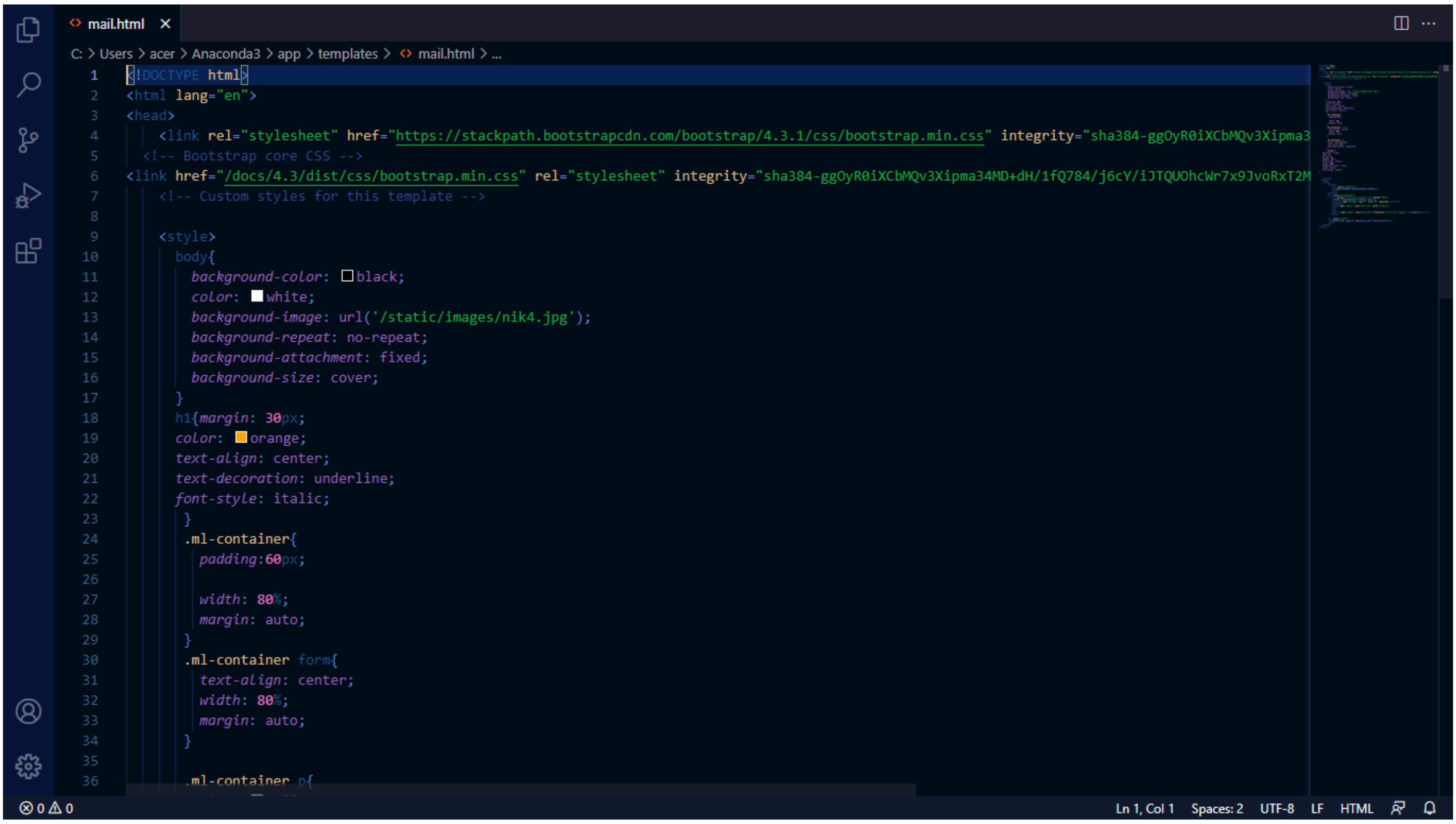
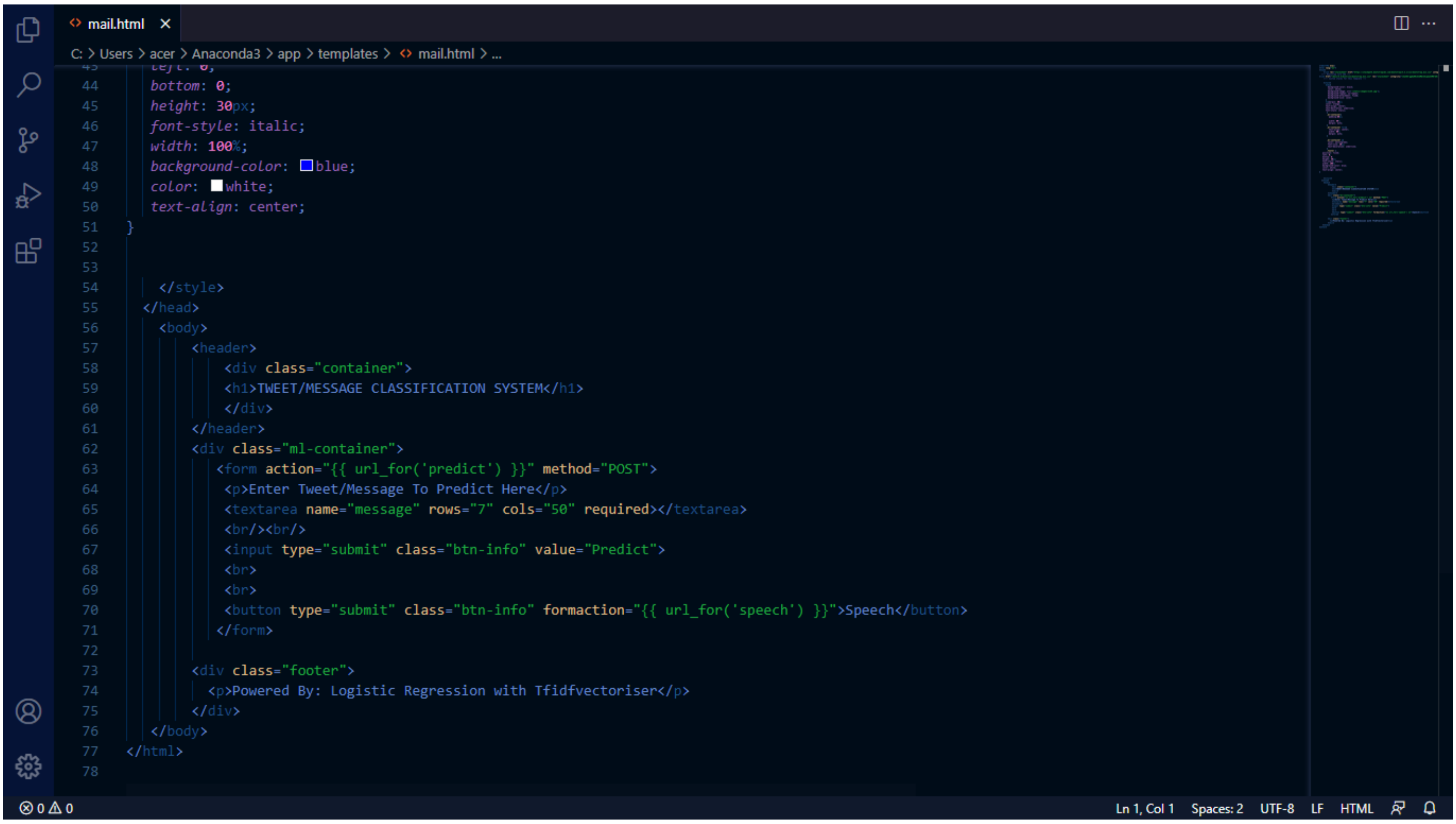

127.0.0.1:5000/mail
127.0.0.1:5000/mail
TWEET/MESSAGE CLASSIFICATION SYSTEM
Enter Tweet/Message To Predict Here
Predict
Speech
Powered By: Logistic Regression with Tfidfvectoriser

Figure 11.3: Checker page

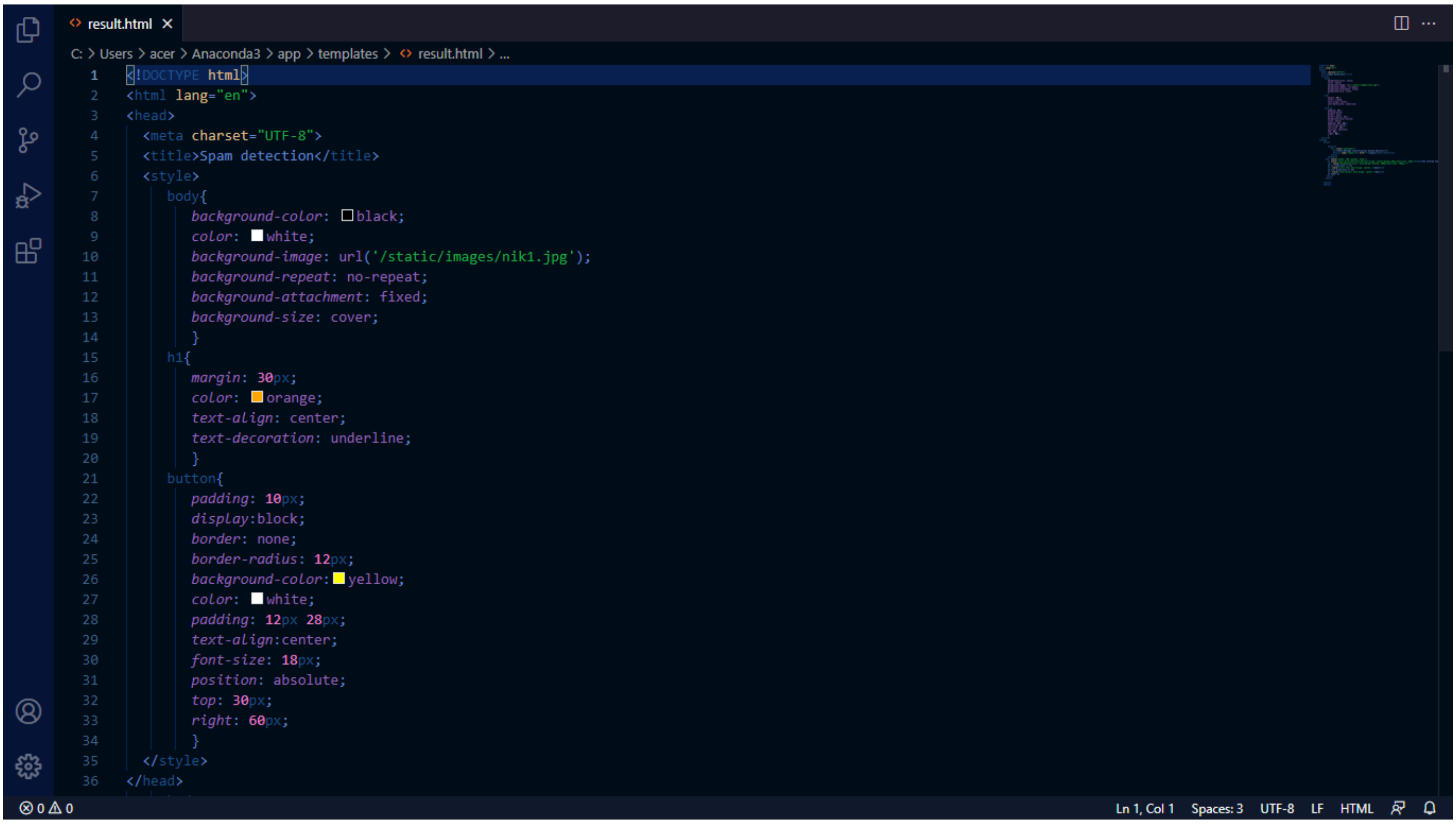

mail.html
C: > Users > acer > Anaconda3 > app > templates > mail.html > ...
Ln 1, Col 1
Spaces: 2
UTF-8
LF
HTML

```
<!DOCTYPE html>
<html lang="en">
<head>
    <link rel="stylesheet" href="https://stackpath.bootstrapcdn.com/bootstrap/4.3.1/css/bootstrap.min.css" integrity="sha384-ggOyR0iXCbMQv3Xipma3
  <!-- Bootstrap core CSS -->
<link href="/docs/4.3/dist/css/bootstrap.min.css" rel="stylesheet" integrity="sha384-ggOyR0iXCbMQv3Xipma34MD+dH/1fQ784/j6cY/iJTQUOhcWr7x9JvoRxT2M
    <!-- Custom styles for this template -->

    <style>
      body{
        background-color: black;
        color: white;
        background-image: url('/static/images/nik4.jpg');
        background-repeat: no-repeat;
        background-attachment: fixed;
        background-size: cover;
      }
      h1{margin: 30px;
      color: orange;
      text-align: center;
      text-decoration: underline;
      font-style: italic;
       }
       .ml-container{
         padding:60px;

         width: 80%;
         margin: auto;
       }
       .ml-container form{
         text-align: center;
         width: 80%;
         margin: auto;
       }

       .ml-container p{
```

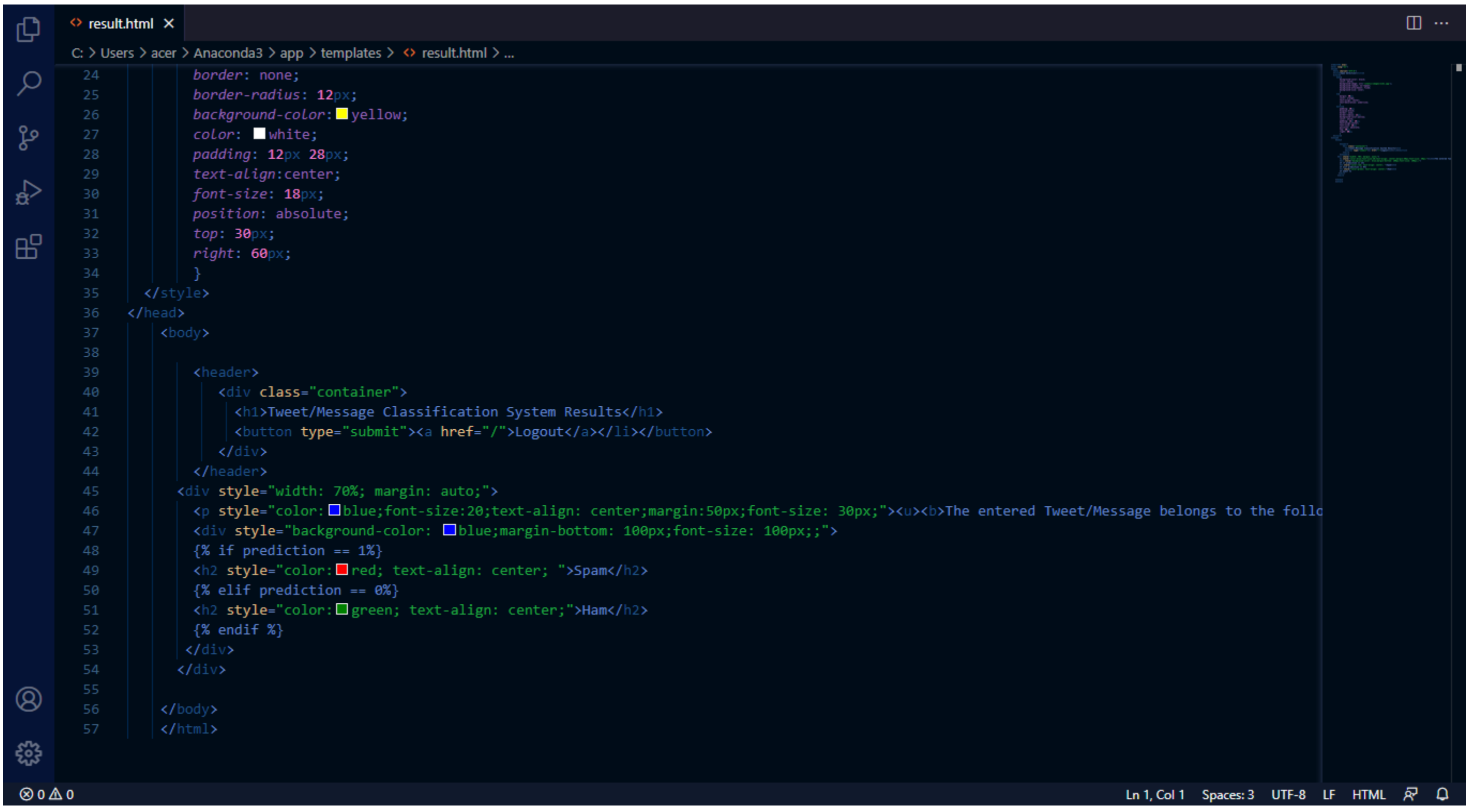

mail.html
C: > Users > acer > Anaconda3 > app > templates > mail.html > ...
⊗ 0 ⚠ 0
Ln 1, Col 1 Spaces: 2 UTF-8 LF HTML

```
   left: 0;
   bottom: 0;
   height: 30px;
   font-style: italic;
   width: 100%;
   background-color: blue;
   color: white;
   text-align: center;
}


    </style>
  </head>
    <body>
        <header>
            <div class="container">
            <h1>TWEET/MESSAGE CLASSIFICATION SYSTEM</h1>
            </div>
        </header>
        <div class="ml-container">
           <form action="{{ url_for('predict') }}" method="POST">
            <p>Enter Tweet/Message To Predict Here</p>
            <textarea name="message" rows="7" cols="50" required></textarea>
            <br/><br/>
            <input type="submit" class="btn-info" value="Predict">
            <br>
            <br>
            <button type="submit" class="btn-info" formaction="{{ url_for('speech') }}">Speech</button>
           </form>

        <div class="footer">
          <p>Powered By: Logistic Regression with Tfidfvectoriser</p>
        </div>
   </body>
</html>
```

result.html


result.html
C: > Users > acer > Anaconda3 > app > templates > result.html > ...
⊗ 0 ⚠ 0
Ln 1, Col 1 Spaces: 3 UTF-8 LF HTML

```
<!DOCTYPE html>
<html lang="en">
<head>
  <meta charset="UTF-8">
  <title>Spam detection</title>
  <style>
     body{
        background-color: black;
        color: white;
        background-image: url('/static/images/nik1.jpg');
        background-repeat: no-repeat;
        background-attachment: fixed;
        background-size: cover;
        }
     h1{
        margin: 30px;
        color: orange;
        text-align: center;
        text-decoration: underline;
        }
     button{
        padding: 10px;
        display:block;
        border: none;
        border-radius: 12px;
        background-color: yellow;
        color: white;
        padding: 12px 28px;
        text-align:center;
        font-size: 18px;
        position: absolute;
        top: 30px;
        right: 60px;
        }
  </style>
</head>
```



result.html
C: > Users > acer > Anaconda3 > app > templates > result.html > ...
0 0
Ln 1, Col 1
Spaces: 3
UTF-8
LF
HTML

```
                border: none;
                border-radius: 12px;
                background-color: yellow;
                color: white;
                padding: 12px 28px;
                text-align:center;
                font-size: 18px;
                position: absolute;
                top: 30px;
                right: 60px;
                }
        </style>
    </head>
        <body>

            <header>
               <div class="container">
                 <h1>Tweet/Message Classification System Results</h1>
                 <button type="submit"><a href="/">Logout</a></li></button>
               </div>
            </header>
          <div style="width: 70%; margin: auto;">
            <p style="color:blue;font-size:20;text-align: center;margin:50px;font-size: 30px;"><u><b>The entered Tweet/Message belongs to the follo
            <div style="background-color: blue;margin-bottom: 100px;font-size: 100px;;">
            {% if prediction == 1%}
            <h2 style="color:red; text-align: center; ">Spam</h2>
            {% elif prediction == 0%}
            <h2 style="color:green; text-align: center;">Ham</h2>
            {% endif %}
           </div>
          </div>

        </body>
        </html>
```

Format - I

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Deemed to be University u/s 3 of UGC Act, 1956)

Office of Controller of Examinations

REPORT FOR PLAGIARISM CHECK ON THE DISSERTATION/PROJECT REPORTS FOR UG/PG PROGRAMMES

(**To be attached in the dissertation/ project report**)

1	Name of the Candidate **(IN BLOCK LETTERS)**	SHIV NOLIYAN
2	Address of the Candidate	32k, Kamboyan Colony, ThanaBhawan, Distt. Shamli, Uttar Pradesh, 247777 **Mobile Number :** 9084684215
3	Registration Number	RA1711003010265
4	Date of Birth	16/01/1999
5	Department	Computer Science And Engineering
6	Faculty	Mr. M.Senthil Raja
7	Title of the Dissertation/Project	Classification Of Tweets/Messages Using Logistic Regression and Tf-idfVectoriser Enabled with Text-to-Speech
8	Whether the above project/dissertation is done by	Individual or group : (Strike whichever is not applicable) a) If the project/ dissertation is done in group, then how many students together completed the project : 2 b) Mention the Name & Register number of other candidates : Nikhil Kumar Singh- RA1711003010265
9	Name and address of the Supervisor / Guide	Mr. M.Senthil Raja senthilraja.ma@ktr.srmuniv.ac.in 9884445605 **Mail ID : Mobile Number :**
10	Name and address of the Co-Supervisor / Co- Guide (if any)	**Mail ID : Mobile Number :**

11	Software Used	TURNIT IN
12	Date of Verification	26/05/2021
13	**Plagiarism Details: (to attach the final report from the software)**	

Chapter	**Title of the Chapter**	**Percentage of similarity index (including self citation)**	**Percentage of similarity index (Excluding self citation)**	**% of plagiarism after excluding Quotes, Bibliography, etc.,**
1	INTRODUCTION	0%	0%	0%
2	LITERATURE SURVEY	1%	1%	1%
3	SYSTEM ARCHITECTURE	0%	0%	0%
4	REQUIREMENT SPECIFICATION	0%	0%	0%
5	SYSTEM DESIGN	0%	0%	0%
6	MODULES	1%	1%	1%
7	BENEFITS	0%	0%	0%
8	RESULT	<1%	<1%	<1%
9	CONCLUSION	0%	0%	0%
10	FUTURE ENHANCEMENTS	<1%	<1%	<1%
Appendices		0%	0%	0%

I / We declare that the above information have been verified and found true to the best of my / our knowledge.

SHIV NOLIYAN **Signature of the Candidate**	**Name & Signature of the Staff (Who uses the plagiarism check software)**
M.Senthil Raja **Name & Signature of the Supervisor/Guide**	**Name & Signature of the Co-Supervisor/Co-Guide**

Name & Signature of the HOD

TFIDF

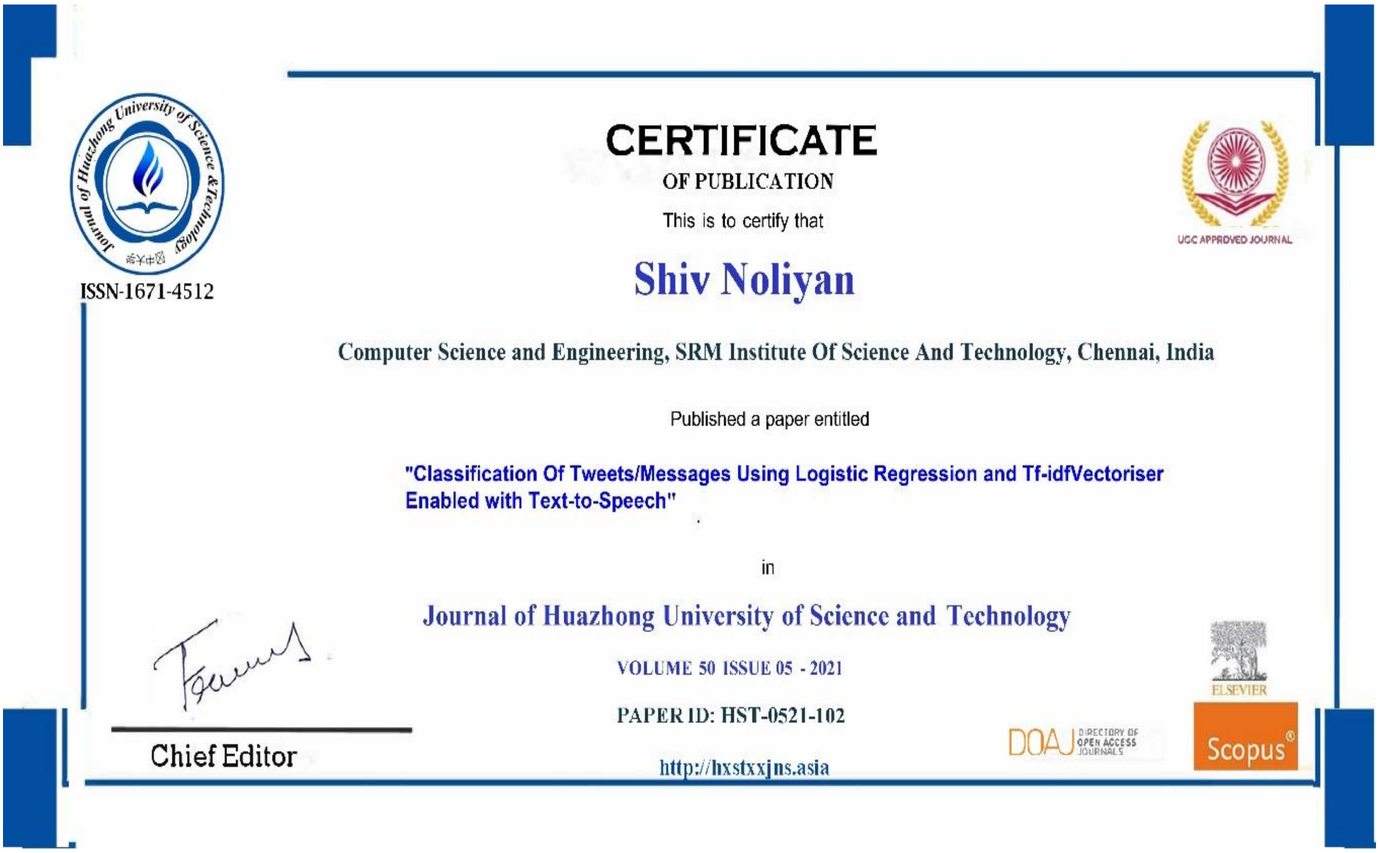
ORIGINALITY REPORT

3%	2%	2%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Bahia Halawi, Azzam Mourad, Hadi Otrok, Ernesto Damiani. "Few are as Good as Many: An Ontology-Based Tweet Spam Detection Approach", IEEE Access, 2018 Publication	1%
2	Submitted to University of Pretoria Student Paper	1%
3	www.scribd.com Internet Source	1%
4	www.litcharts.com Internet Source	<1%
5	web.ma.utexas.edu Internet Source	<1%
6	www.boute.ir Internet Source	<1%

Exclude quotes	On	Exclude matches	< 3 words
Exclude bibliography	On		



ISSN-1671-4512

CERTIFICATE

OF PUBLICATION

This is to certify that

UGC APPROVED JOURNAL

Shiv Noliyan

Computer Science and Engineering, SRM Institute Of Science And Technology, Chennai, India

Published a paper entitled

"Classification Of Tweets/Messages Using Logistic Regression and Tf-idfVectoriser Enabled with Text-to-Speech"

in

Journal of Huazhong University of Science and Technology

VOLUME 50 ISSUE 05 - 2021

PAPER ID: HST-0521-102

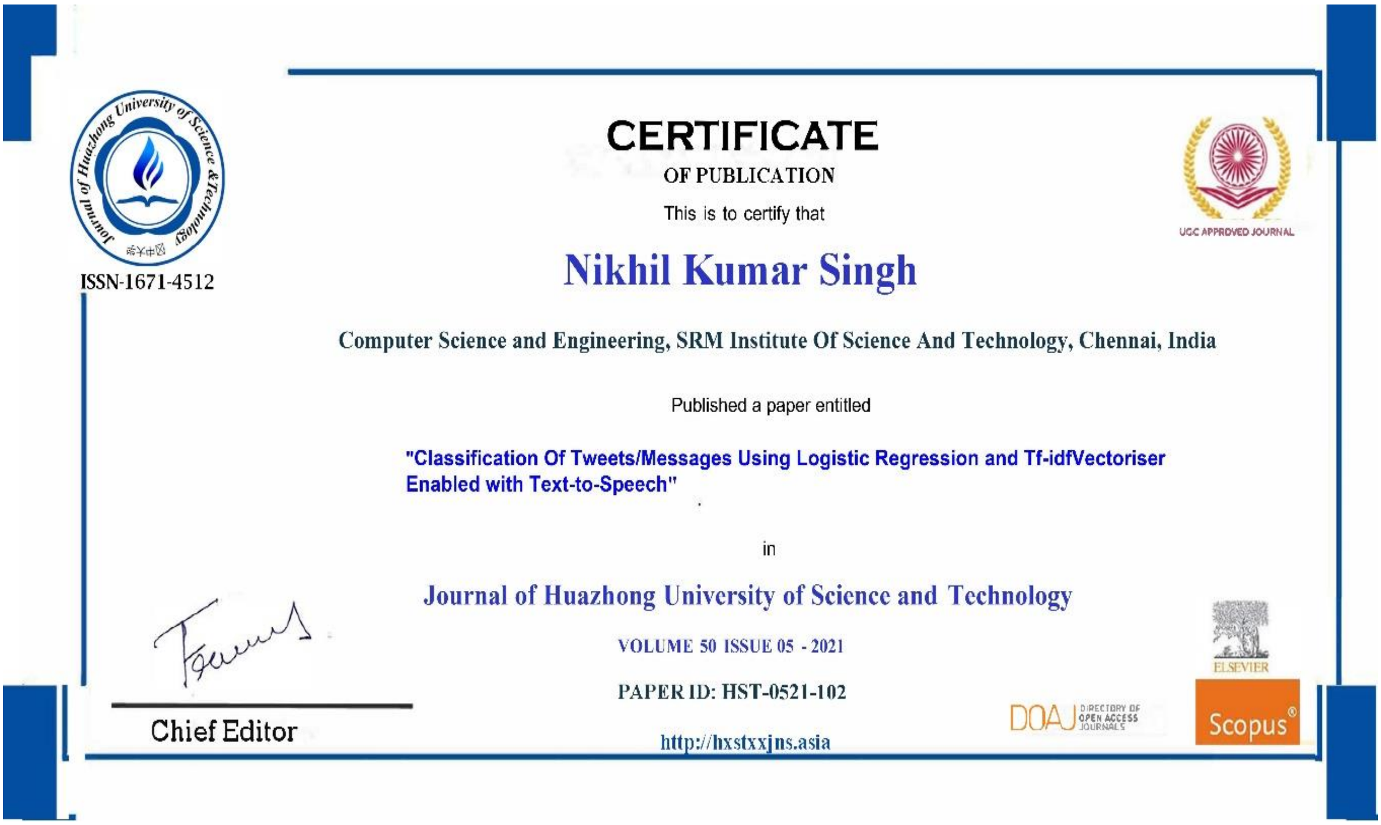
Chief Editor

http://hxstxxjns.asia

DOAJ DIRECTORY OF OPEN ACCESS JOURNALS

ELSEVIER

Scopus®



ISSN-1671-4512

CERTIFICATE

OF PUBLICATION

This is to certify that

UGC APPROVED JOURNAL

Nikhil Kumar Singh

Computer Science and Engineering, SRM Institute Of Science And Technology, Chennai, India

Published a paper entitled

"Classification Of Tweets/Messages Using Logistic Regression and Tf-idfVectoriser Enabled with Text-to-Speech"

in

Journal of Huazhong University of Science and Technology

VOLUME 50 ISSUE 05 - 2021

PAPER ID: HST-0521-102

Chief Editor

http://hxstxxjns.asia

DOAJ DIRECTORY OF OPEN ACCESS JOURNALS

ELSEVIER

Scopus®