



CALIFORNIA STATE UNIVERSITY  
**FULLERTON**<sup>TM</sup>

## Department of Computer Science

This project has been satisfactorily demonstrated and is of suitable form.

This project report is acceptable in partial completion of the requirements for the Master of Science degree in Computer Science.

**DATA MINING IN PUBLIC TRANSPORTATION SAFETY**

---

Project Title (type)

**PHUC CONG LE**

---

Student Name (type)

**Dr. ANAND V. PANANGADAN**

---

Advisor's Name (type)

---

Advisor's signature

Date

**Dr. LIDIA MORRISON**

---

Reviewer's name

---

Reviewer's signature

Date

## **Abstract**

Nowadays, public transportation is intentionally developed to encourage more and more riders. However, many people are concerning about the safety when commuting by public transit. To understand more about the concerns, the public transit authority may individually read the opinions of the riders by making surveys; however, that may take excess time and effort to achieve the goal. A better approach is to automatically analyze the tweets from Twitter using machine learning algorithms and data mining techniques. This approach takes advantage of increasing willingness to share ideas from users on social media. In summary, I'm going to develop a project based on the concept of sentiment analysis with the input from a broad number of tweets, but instead of positive or negative sentiment output, the goal of my project is to identify the safety or not-safety signal associated with each tweet.

## **Table of Content**

1. Introduction	1
1.a. Description of the Problem	1
1.b. Project Objectives	1
1.c. Development Environment	2
1.d. Operational Environment	2
2. Requirements Description	3
3. Design Description	5
4. Implementation	9
5. Test and Integration (plan and results)	13
6. Installation and Operating Instructions	13
7. Recommendation for Enhancement	14
8. Bibliography	15
9. Source Code	15

## **1. Introduction**

### **1.a. Description of the Problem**

Nowadays, public transportation is intentionally developed to encourage more and more riders. However, many people are concerning about the safety when commuting by public transit. To understand more about the concerns, the public transit authority may individually read the opinions of the riders by making surveys; however, that may take excess time and effort to achieve the goal. A better approach is to automatically analyze the tweets from Twitter using machine learning algorithms and data mining techniques. This approach takes advantage of increasing willingness to share ideas from users on social media. In summary, I'm going to develop a project based on the concept of sentiment analysis with the input from a broad number of tweets, but instead of positive or negative sentiment output, the goal of my project is to identify the safety or not-safety signal associated with each tweet.

### **1.b. Project Objectives**

The objective of my project is to apply machine learning and data mining techniques in public transportation safety. The goal of this project is to build up a web-based application that analyzes a large amount of data about the public transit from Twitter, and the outcome would be the indication of safety levels of each tweet like safety, unsafety, neutral, or unknown (not related). From the result of the analysis, there will be some additional statistic information shown on the interface like how many tweets of the total mentions containing safety indication, the presentation in pie chart, and the map associated with the tweets' geolocation.

This system is significant enough for a master's project because it plays a role as a collection of the skill and knowledge that I have accumulated, learned and discovered at CSU Fullerton. At the completion, my project can be a practical use in the public safety sector with the capability of transforming the massive volume of unorganized data from online social media into useful information that may support decision-making, crime prevention, and threat warning identification.

### 1.c. Development Environment

The application is implemented using Visual Studio Code integrated development environment. It can be downloaded from <https://code.visualstudio.com>.

The software can be debugged using Google Chrome and the built-in system Terminal.

- Programming language: Python 3.6.7, JavaScript
- Framework: Django 2.0.3
- Libraries: Bootstrap, jQuery, Tweepy, Uszipcode
- Database: SQLite 2.6.0
- Others: CSS3, html5, Twitter API, NLTK, TensorFlow, NumPy, OpenLayers API for map view, Google Charts

### 1.d. Operational Environment

The software should be run if all of its libraries and dependencies are installed properly into a computer running Ubuntu 17.04 or newer.



## **2. Requirements Description**

- The system shall allow the user to download tweets by zip code.
- The application shall display all the tweets along with the prediction result for each tweet.
- The website shall display the map showing the prediction result in different color bookmarks according to the safety level of each tweet.
- The application shall generate the pie chart representing outcome proportions.
- The website will have a user-friendly interface.
- The website should have a navigation bar for all of the web pages.
- The website shall display a progress bar icon while downloading the tweets.
- The result page shall be capable of grouping tweets by the safety level.
- The system shall be able to export the report csv file from the analysis.

## Data Mining in Public Transportation Safety

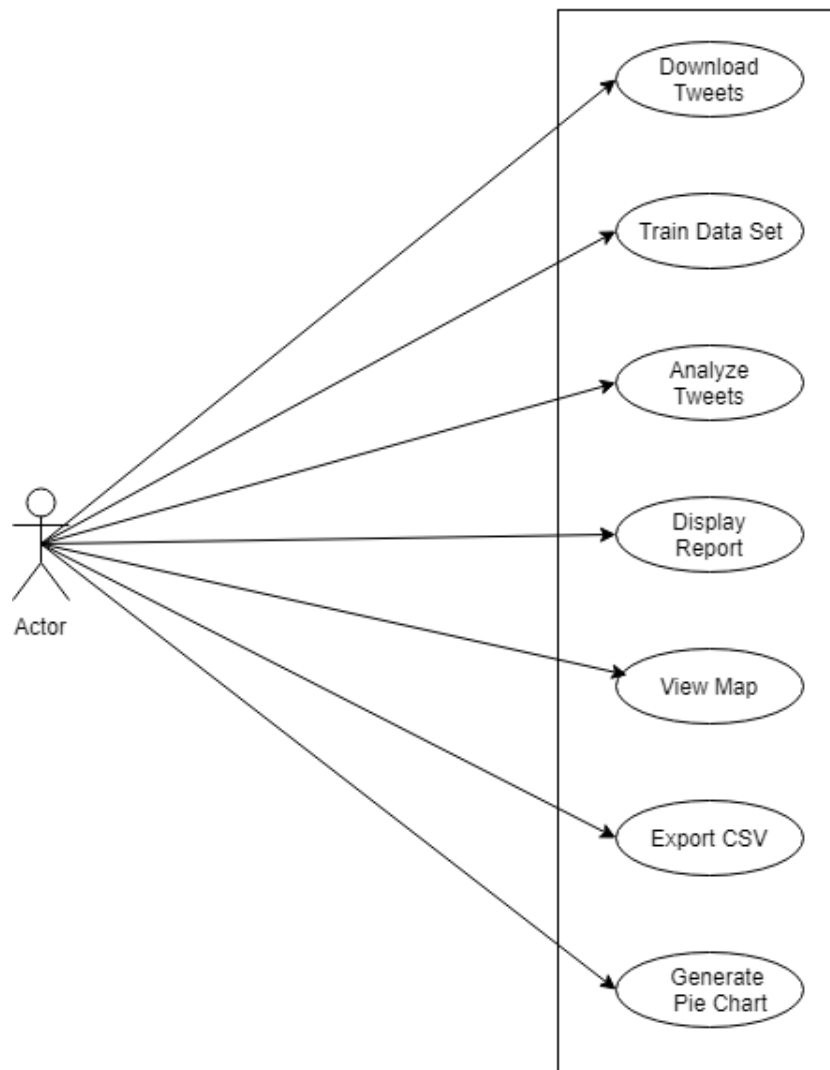


Figure 2.1 Use Case Diagram for Data Mining in Public Transportation Safety

### 3. Design Description

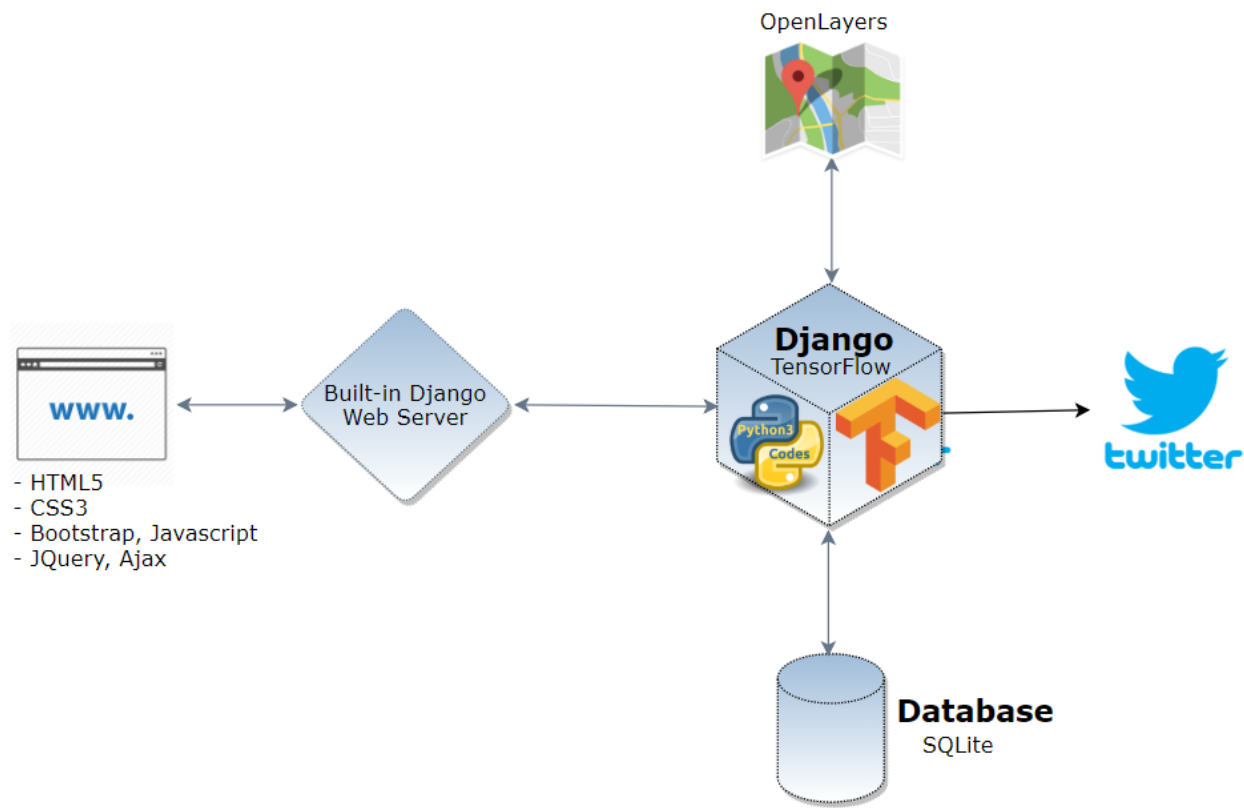


Figure 3.1 – The architecture

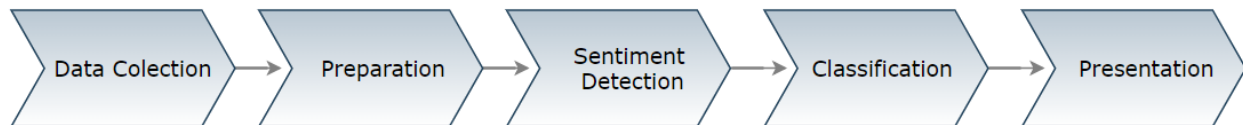


Figure 3.2 – Data processing approach

The figure 3.2 state diagrams will visually explain the steps for data processing. In general, a sentiment analysis usually has to go through five steps in processing the unstructured data into the more meaningful outputs; they are the collection of data, preparation, sentiment detection, classification, and presentation of the output. In order to improve the accuracy of sentiment analysis, the data preparation step should be intensively undertaken for a better classification as well as the outputs. Today, most of the data is retrieved from Twitter because it is a high-volume social media platform where the users tend to actively and more willingly express their opinions and interests and the information is rapidly

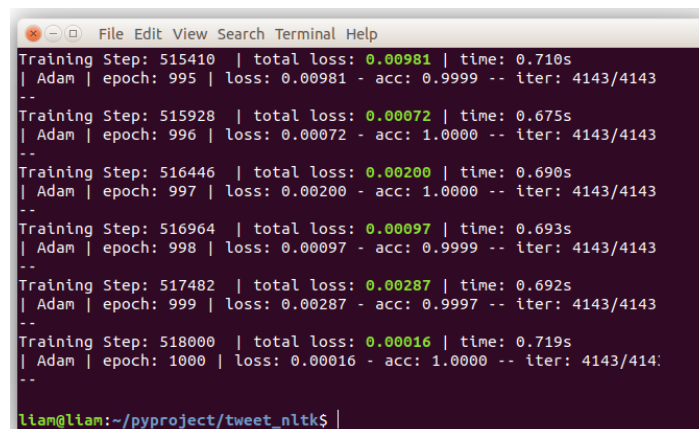


widespread (Zimbra, Abbasi, Zeng, & Chen, 2018; Fortuna & Nunes, 2018). For better data preparation, the data should be polished and filtered. The details can be found in the code comments below.

```
def preprocess(tweet):  
  
    # Convert to lower case  
    tweet = tweet.lower()  
  
    # Replace links with the word URL  
    tweet = re.sub('((www\.[^\s]+)|(https?://[^\s]+))', 'URL', tweet)  
  
    # Replace @username with ""  
    tweet = re.sub('@[^\s]+', '', tweet)  
  
    # Replace #word with word  
    tweet = re.sub(r'#([^\s]+)', r'\1', tweet)  
  
    # Removed punctuation using regex  
    tweet = re.sub(r'[^\w\s]', '', tweet)  
  
    # Removed non-english words  
    tweet = " ".join(w for w in nltk.wordpunct_tokenize(tweet) if w.lower() in words or not w.isalpha())  
  
    #remove unicodes  
    tweet = tweet.encode('ascii', errors='ignore').strip().decode('ascii')  
  
    # Removed stopwords  
    tweet = " ".join(w for w in nltk.wordpunct_tokenize(tweet) if not w in stop_words)  
  
    #stemming of words  
    tweet = " ".join(porter.stem(word) for word in nltk.wordpunct_tokenize(tweet))  
  
    return tweet
```

Figure 3.3 – preprocess each Tweet

In this project, my dataset includes approximately 6,600 tweets for training purpose, and it took about 24 minutes for finish building the model based on those tweets. The figure 3.4 shows the training process.



```
File Edit View Search Terminal Help  
Training Step: 515410 | total loss: 0.00981 | time: 0.710s  
| Adam | epoch: 995 | loss: 0.00981 - acc: 0.9999 -- iter: 4143/4143  
--  
Training Step: 515928 | total loss: 0.00072 | time: 0.675s  
| Adam | epoch: 996 | loss: 0.00072 - acc: 1.0000 -- iter: 4143/4143  
--  
Training Step: 516446 | total loss: 0.00200 | time: 0.690s  
| Adam | epoch: 997 | loss: 0.00200 - acc: 1.0000 -- iter: 4143/4143  
--  
Training Step: 516964 | total loss: 0.00097 | time: 0.693s  
| Adam | epoch: 998 | loss: 0.00097 - acc: 0.9999 -- iter: 4143/4143  
--  
Training Step: 517482 | total loss: 0.00287 | time: 0.692s  
| Adam | epoch: 999 | loss: 0.00287 - acc: 0.9997 -- iter: 4143/4143  
--  
Training Step: 518000 | total loss: 0.00016 | time: 0.719s  
| Adam | epoch: 1000 | loss: 0.00016 - acc: 1.0000 -- iter: 4143/4143  
--  
liam@liam:~/pyproject/tweet_nltk$
```

Figure 3.4 – Dataset trained

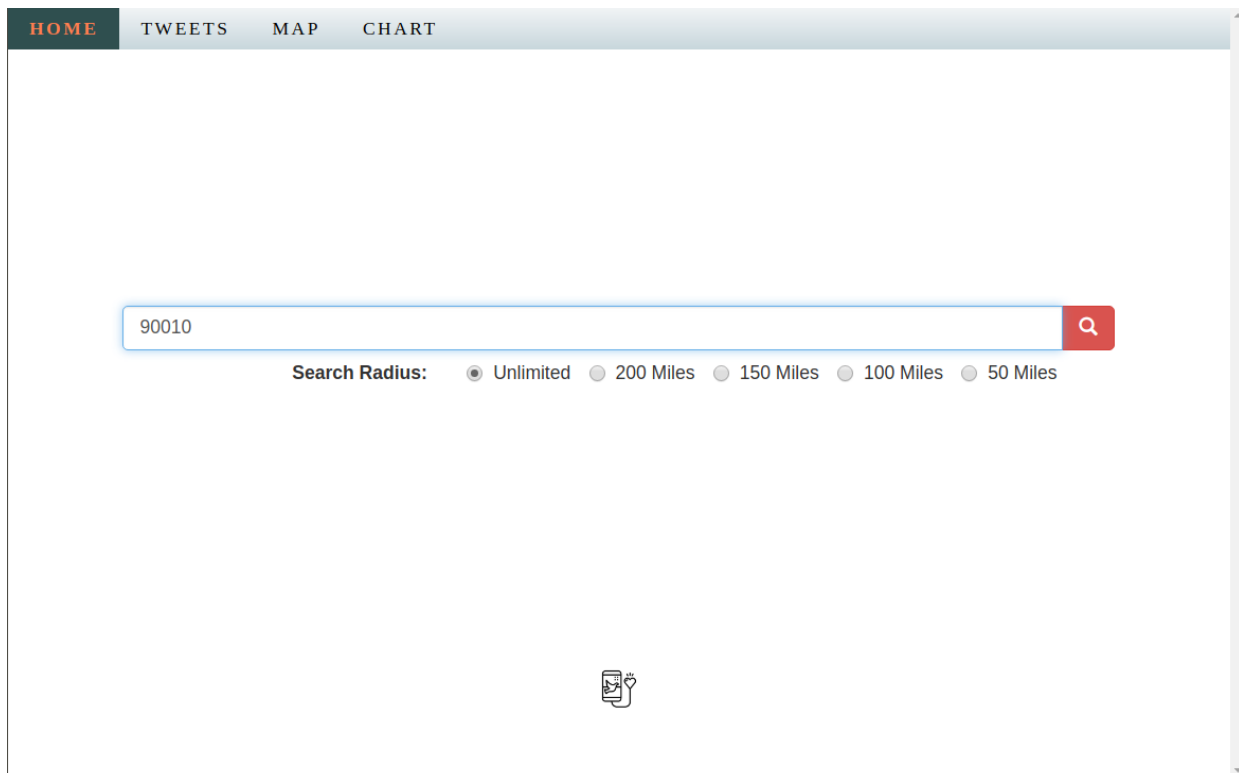


Figure 3.5 – Home page screen captured

HOME	TWEETS	MAP	CHART
Neutral	Oh wonderful, there is a service, <a href="https://t.co/v3ovkcynBQ">https://t.co/v3ovkcynBQ</a> , for buying train tickets for long cross-continent journeys in Europe		
Neutral	im already crying about the nct concert and black hair jimin and the puppy interview shit and my grades and my frie... <a href="https://t.co/qilvBpfd7r">https://t.co/qilvBpfd7r</a>		
Unknown	Thotiana is an environmentally conscious activist who utilizes public transport. We should all be like Thotiana and indeed, bus it down. 🚌		
Unsafe	@JoeBugBuster A6. In Arizona the worst is by bus. Itd take hours to go 20 miles.. but if its 117° out it prolly bea... <a href="https://t.co/G0jovPSqy1">https://t.co/G0jovPSqy1</a>		
Neutral	The Ride On route 100 bus provides all-day express service (via I-270) from the Germantown Transit Center to Shady... <a href="https://t.co/7vQn3N59Si">https://t.co/7vQn3N59Si</a>		
Neutral	@jimbobbennett S'up in the Northgate Mall. About 15-20 minutes by bus from downtown.		
Unknown	@Kos2Order @The_Trump_Train @realDonaldTrump I agree! Let's look st impeachment.		
Neutral	After today I think the flying Scotsman need not run on any part of the British mainline rail network for ever. Yet... <a href="https://t.co/Tu0zZkz7yt">https://t.co/Tu0zZkz7yt</a>		
Unknown	that nct issue with the creepy girls on their bus is actually terrifying. people like that shouldn't be allowed to... <a href="https://t.co/UZ2q5b8kOk">https://t.co/UZ2q5b8kOk</a>		
Neutral	I've been saying something is wrong with her for months now! She seems to loose her train of thought when speaking!... <a href="https://t.co/RwRnetkeVP">https://t.co/RwRnetkeVP</a>		
Neutral	I'm being very serious when I say this. By the year 2024 @RealCandaceO will be old enough to run for President. T... <a href="https://t.co/G5SzpelQOP">https://t.co/G5SzpelQOP</a>		
Unknown	nah wtf don't fucking go in their bus even if you're given the option don't do it????? that just,, and making th... <a href="https://t.co/AtQJ2UyZ7Z">https://t.co/AtQJ2UyZ7Z</a>		
Unknown	Super fun on stream today! Gave away like 500K or more gold away ☺ Starting Movie Night event right now- How to Tra... <a href="https://t.co/xNTBNKUPAi">https://t.co/xNTBNKUPAi</a>		
Unknown	@PennySc05759227 @TimHunt78506372 @American_JimG @The_Trump_Train @realDonaldTrump Well his stuff will have a tariff then.		
Unknown	@The_Trump_Train @realDonaldTrump Trump supporters still be like <a href="https://t.co/CWhXmaz4Jt">https://t.co/CWhXmaz4Jt</a>		
Unknown	Guys if you witnessed what I just witnessed.. So I'm at the bus stop right now on my way home from work and there's... <a href="https://t.co/KFZcXmhajl">https://t.co/KFZcXmhajl</a>		
Unknown	While going for our bath post muddy dog park, we found this awesome pamphlet about #DogGuides, and raising money to... <a href="https://t.co/WBMErZopLx">https://t.co/WBMErZopLx</a>		
Unknown	@PencesAngryEyes @realDonaldTrump Or an Israeli train station. Assuming, of course, that Trump's still-incomplete "... <a href="https://t.co/a4TaXycdDx">https://t.co/a4TaXycdDx</a>		
Unknown	Talked to drivers and monitors after a productive meeting at Colonial HS. Thank you for your work, CF School Bus D... <a href="https://t.co/HVzPBsZuK1">https://t.co/HVzPBsZuK1</a>		
Unknown	@bjeffcoat9 @richhomiecait no you just suck .. we literally playing train and you put out skip blue 4 and 1 talking bout uno uno out 🤔🤔🤔		
Unknown	This is a big development for #mobility and it's great to see my old partners at RTD leading the way. A big step fo... <a href="https://t.co/cIMZhPcwJs">https://t.co/cIMZhPcwJs</a>		

Figure 3.6 – Analysis result in list view

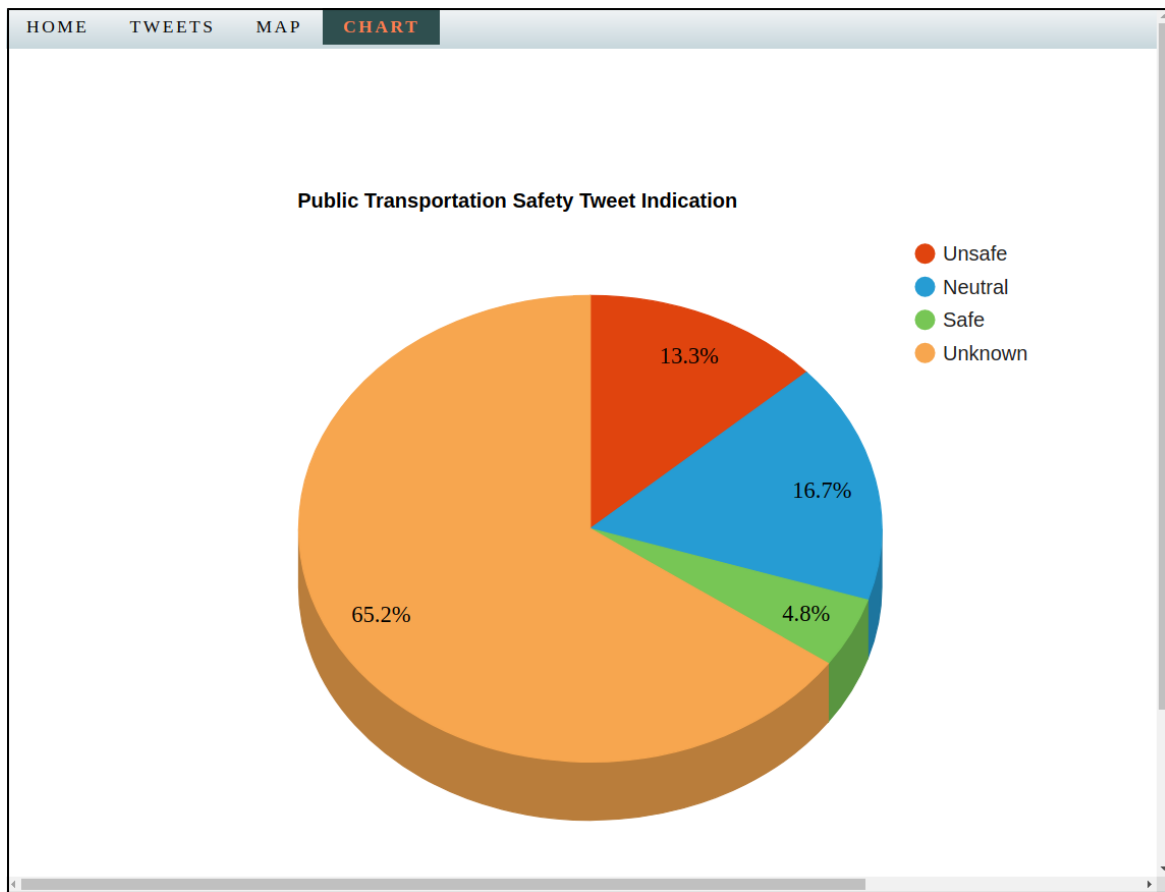


Figure 3.7 – Analysis result in pie chart

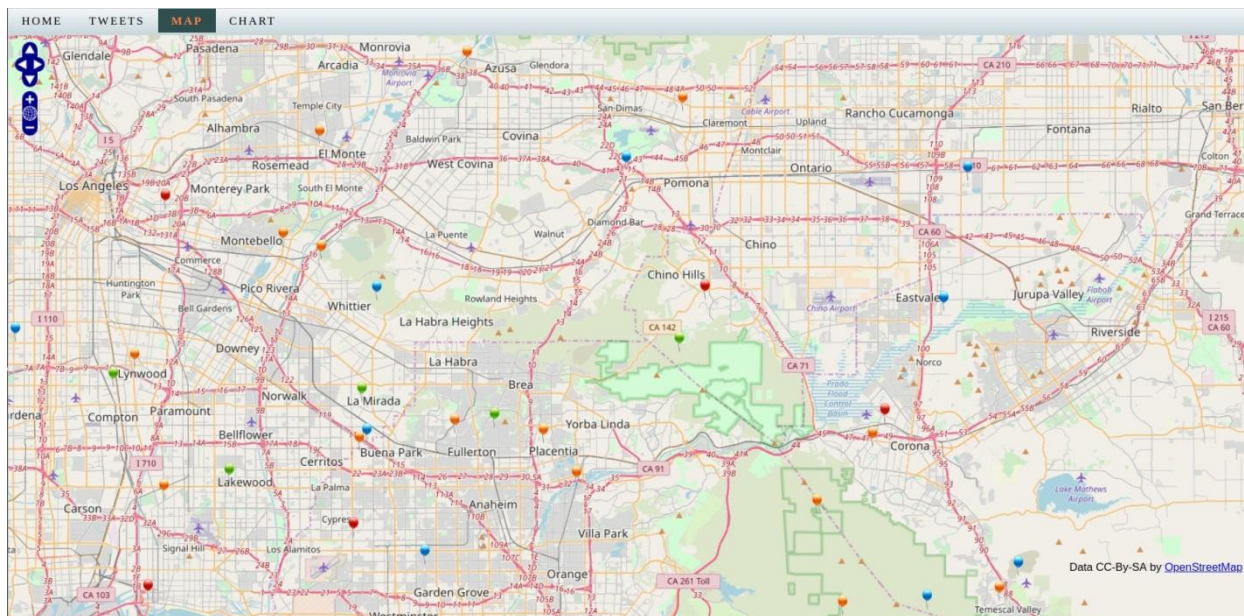


Figure 3.8 – Analysis result in map view with bookmarks

## 4. Implementation

This project has two sections. The first one is coded to train the data which given a collection of tweets with predefined labels and output the model. The second one is the main section that will collect the tweets, conduct the analysis based on the built model, and display the analysis result in a list view, graph, and map.

### 4.1 Training the Dataset

There are 2 files named `classify.py` and `data.json` inside folder “training.” In order to train the dataset, open Terminal, change directory to “training” directory and type the following command: ***python3 classify.py***. After the command is run successfully, there are five (5) files and a folder will be produced, and they present the model outcome.

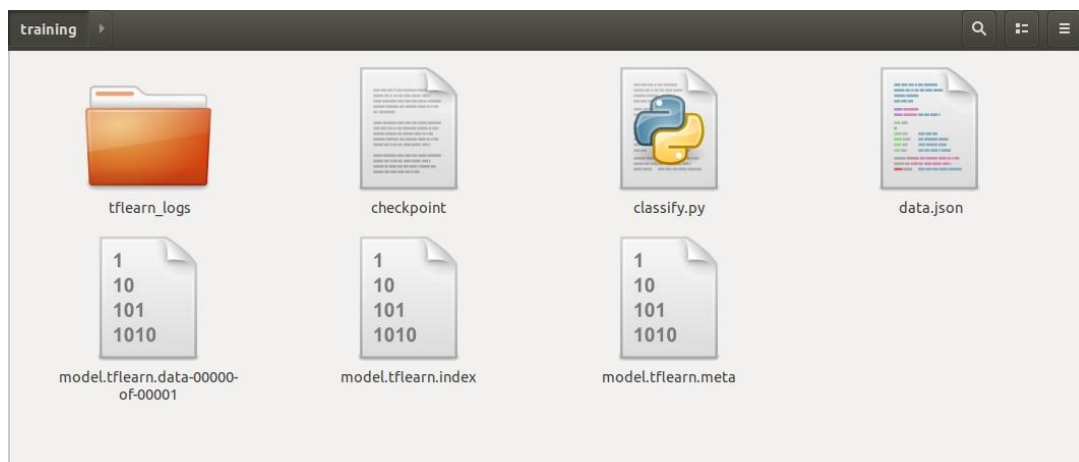


Figure 4.1.1 – Training dataset file structure

```

1  {
2    "pos": [
3      "Dogs help secure Araneta bus terminal",
4      "train station is well managed",
5      "Accident cleared in on Rittiman Rd near Seguin Rd traffic",
6      "professional bus driver",
7      "have a safe bus trip to Los Angeles",
8      "bus service resume after fatal crash",
9      "ANY POTENTIAL SAFETY TWEETS THAT ARE RELATED TO PUBLIC TRANSPORTATION SAFTTY TOPIC"
10   ],
11   "neg": [
12     "deer staying on the railroad",
13     "animal crossing the bus road",
14     "11 killed in bus accident in",
15     "everyone smokes on the bus",
16     "winter torm, stuck inside the bus",
17     "Bus Slams Into store front",
18     "ANY POTENTIAL UNSAFETY TWEETS THAT ARE RELATED TO PUBLIC TRANSPORTATION SAFTTY TOPIC"
19   ],
20   "neu": [
21     "It was night bus or Shinkansen so",
22     "More so if I can get on and off the track before the train comes",
23     "what bus was it",
24     "i my bus stop",
25     "And off goes the train horn good night shirt",
26     "ANY NEUTRAL TWEETS THAT ARE RELATED TO PUBLIC TRANSPORTATION SAFTTY TOPIC"
27   ],
28   "unk": [
29     "That a boy A Train",
30     "The_Trump_Train",
31     "ThankGodForTrump",
32     "grand funk railroad",
33     "At the end of the day the kid is a... ",
34     "I missed my train Bitcoin never misses a block ",
35     "i LIKe COFFEE",
36     "train my boy",
37     "Tiger wood won the golf cup",
38     "ANY TWEETS THAT ARE NOT RELATED TO PUBLIC TRANSPORTATION SAFTTY TOPIC"
39   ]
40 }

```

Figure 4.1.2 – The structure of dataset stored in “data.json”

The figure 4.1.2 shows how the dataset organizes the tweets according to the labels. The real “data.json” is store in the “training” folder that contains over 6,600 tweets for training purpose.

It takes a lot of time and effort to prepare the dataset. With the structure showed in figure 4.1.2. The training process may not be successful if there is even only a strange character like “ (quote) or bytes representing emoji or special character like

\xF0\x9F\x98\x81. I have to remove them manually or search and replace using programming technique. Please check out the Java project name “TextProcessing” to see how I check and refine the raw tweets.

## 4.2 The Main Section of the Project

This section is a Django project named “cs597;” inside this folder, there are 2 main subfolders: “cs597” contains the setting and the configuration for the project. “app” directory is the main structure for the web application.

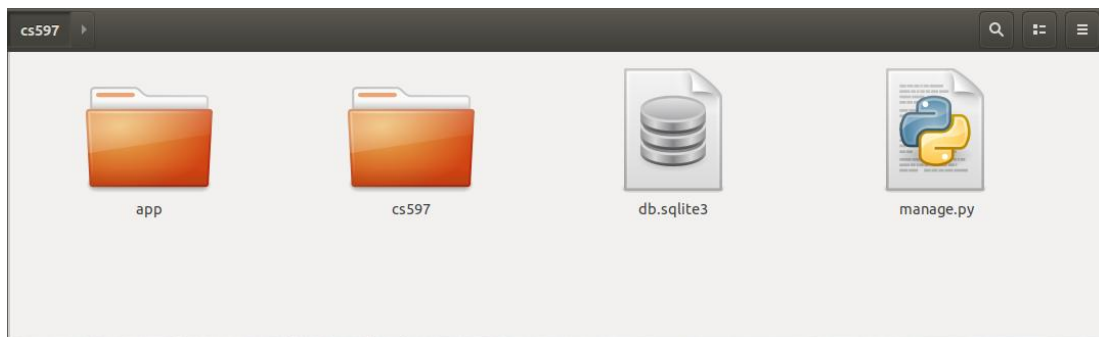


Figure 4.3.1 – File structure of the main section of the project

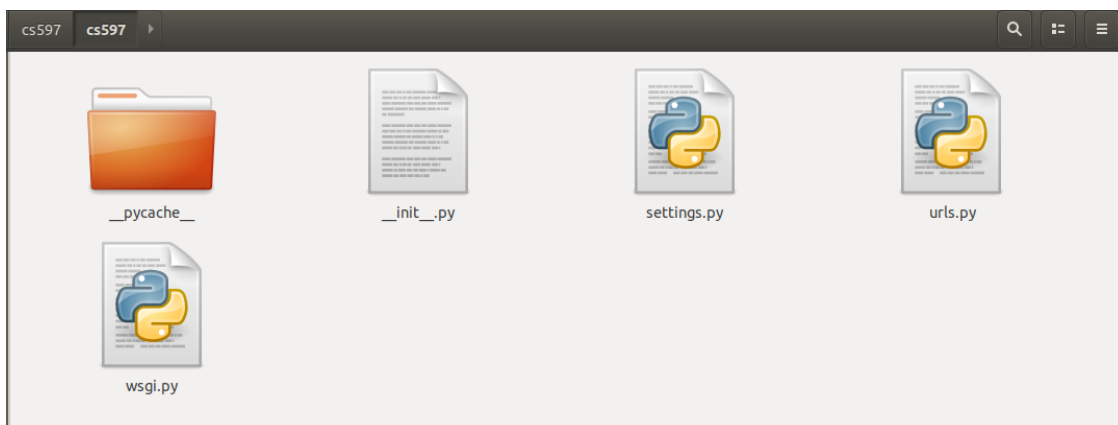


Figure 4.3.2 – File structure of cs597\cs597 directory

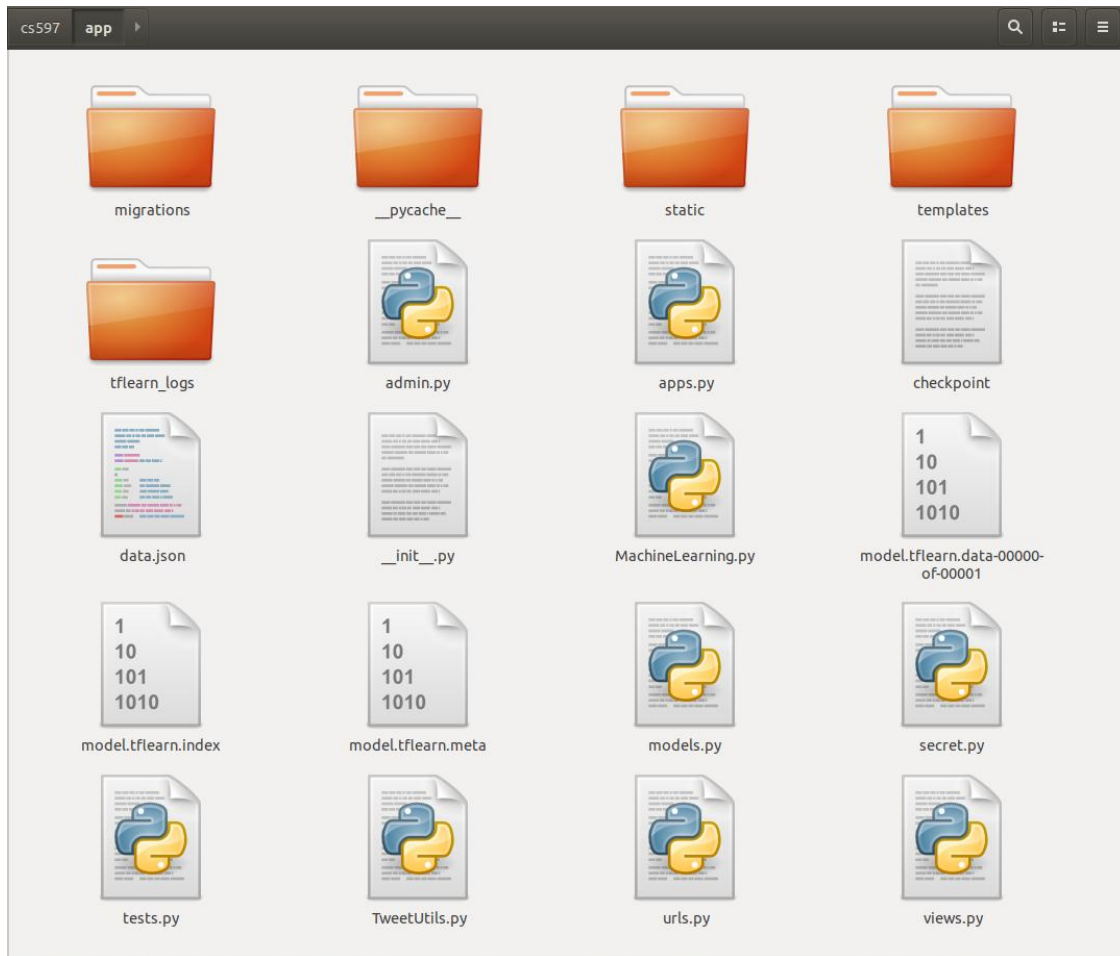


Figure 4.3.3 – File structure of cs597\app directory

File/Directory Name	Description
static	Containing cascading style sheet files, images, icons, and other static assets
templates	Containing all template files in HTML format that are loaded by views for presentation
views.py	Linking the view into the URLs, handling the requests and responses, and providing some useful internal functions.
models.py	Defining some object models that will be mapped to the database table

urls.py	Containing all URL mapping
MachineLearning.py	Providing functions to predict the tweet safety level based on the pre-built model

Table 4.1 – Reference list of main files (see figure 4.2.2)

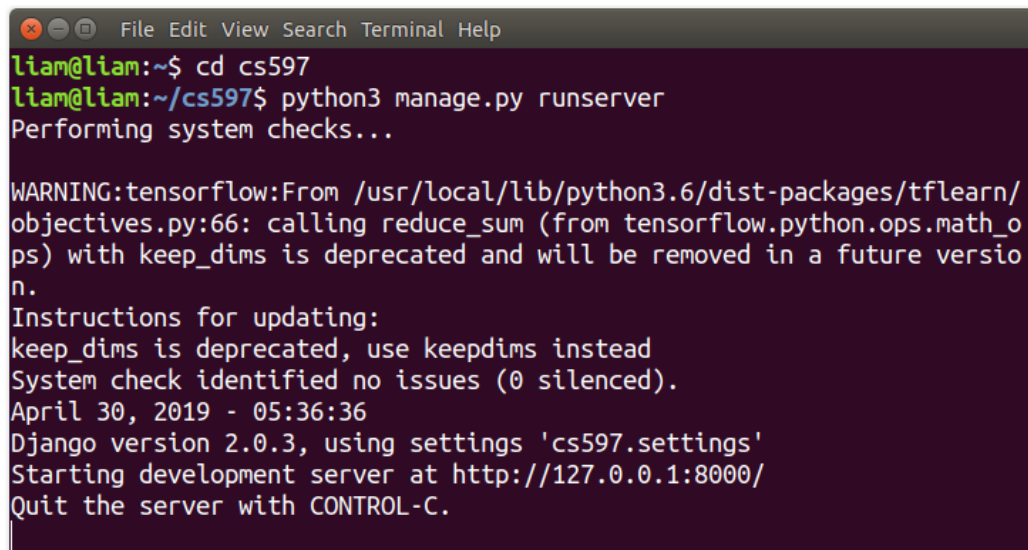
## 5. Test and Integration

- Code Review: Without a peer reviewer, I have to regularly review the code to improve or consolidate the added/alterd/removed functionalities.
- Black-Box Testing: testing the behavior of the application without knowing the internal structure and code of the application. In order to check the input and output, I design a page at **<http://127.0.0.1:8000/predict/>** to check the prediction function. There, the user can input a tweet, submit the form, and see the prediction result displayed on the page.
- Integration Testing: using the bottom-up strategy to make sure each component at a lower levels is tested with components at higher levels until all models are tested. This test is conducted every time the new model is added to the application to make sure that all of the components properly working together.

## 6. Installation and Operating Instructions

- Copy cs597 directory to ***“Home”***
- Open **Terminal**
- Change directory to cs597 by typing: ***cd cs597***
- Run the server by typing: ***python3 manage.py runserver*** as showed in figure 6.1



A terminal window with a dark background and light-colored text. The window title bar shows 'File Edit View Search Terminal Help'. The command prompt is 'liam@liam:~\$'. The user enters 'cd cs597'. The prompt changes to 'liam@liam:~/cs597\$'. The user enters 'python3 manage.py runserver'. The output shows 'Performing system checks...' followed by a warning from TensorFlow about the deprecated 'keep\_dims' parameter. It then provides instructions for updating to 'keepdims'. The system check identifies no issues. The timestamp is 'April 30, 2019 - 05:36:36'. It shows 'Django version 2.0.3, using settings \'cs597.settings\''. The server is starting at 'http://127.0.0.1:8000/'. The instruction to quit the server with 'CONTROL-C.' is shown.

```
liam@liam:~$ cd cs597
liam@liam:~/cs597$ python3 manage.py runserver
Performing system checks...

WARNING:tensorflow:From /usr/local/lib/python3.6/dist-packages/tflearn/
objectives.py:66: calling reduce_sum (from tensorflow.python.ops.math_o
ps) with keep_dims is deprecated and will be removed in a future versio
n.
Instructions for updating:
keep_dims is deprecated, use keepdims instead
System check identified no issues (0 silenced).
April 30, 2019 - 05:36:36
Django version 2.0.3, using settings 'cs597.settings'
Starting development server at http://127.0.0.1:8000/
Quit the server with CONTROL-C.
```

Figure 4.1 How to run the server

- Open Google Chrome and access the home page at the following link:

<http://127.0.0.1:8000/home/>

- At the home page, enter a zipcode, and click on the “**Search**” button.

## 7. Recommendation for Enhancement

- The user may implement a function to calculate the accuracy of the prediction. At the analysis report page, there will have a column to show the prediction’s accuracy percentage to the right of the current table (see figure 3.6)

- For better prediction outcome, the dataset should be increased in size and quality.

- As mentioned before, it takes about 24 minutes to train a dataset containing approximately 6,600 tweets. When the dataset has more tweets added into it, It would be more efficient to train those only those new tweets to have a new model that later consolidated into the current model. In this way, the user doesn’t have to re-train the entire dataset (containing 6,600 tweets + number of new tweets added to the current dataset).

- The application may change the method to download tweets due to the currently limited capability of the library used.

## 8. Bibliography

Zimbra, D., Abbasi, A., Zeng, D., & Chen, H. (2018). The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation. *ACM Trans. Manage. Inf. Syst.* 9, 2, Article 5. Retrieved from <https://doi-org.lib-proxy.fullerton.edu/10.1145/3185045>

Chilukuri, C. (2016). *Identification of safety issues by classification of Twitter data*. Unpublished manuscript, California State University Fullerton

## 9. Source Code

