ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Exploiting Social Media Data for Traffic Monitoring Using the Techniques of Data Mining

Shaikh Kamran , Musaib Shaikh , Alefiya Naseem , Priyanka Kamble

B. E Student, Dept. of Computer Engineering, Trinity College of Engineering and Research, Pune, India

ABSTRACT: Traffic congestion in many cities around the world is severe . The reason is that society has in general become more mobile and this means more people are prepared to commute to work by vehicle than they were before. To tackle these increasing congestion problems we have come up with certain techniques to point out the traffic congested areas. These techniques prove out to be helpful to the pedestrians and tourists who often use roads as their means of transport. Monitoring traffic is a costly and time consuming method that is why we have proposed a user-friendly proficiency to monitor social media(twitter) updates that are updated by the day to day social media users and helps as a tool for keeping the traffic updates up to date.

KEYWORDS: Data Science, Traffic Analysis, Social Media Data, Sentiment Analysis, Statistical models, Dashboard, Visualization, Real-Time.

I. INTRODUCTION

The world population is ever on the increase with the major population residing in the urban areas and metropolitan cities. The number of vehicles on the road is ever increasing and since the last five years it's growing at a rate of 10.16 percent annually. Since the last five years, the number of vehicles on the road is growing. Traffic jams have become a part and parcel of life people living in these metropolitan cities. It not only leads to wastage of fossil fuels which is a non-renewable source of energy but also wastage of time and money. Spending hours in traffic may lead to health hazards and the burning of fuel may cause environmental hazards.

There are a number of methods or approaches which are available or suggested to address traffic problems. The primary approach is, better infrastructure to be constructed like wider roads, bypasses and expressways. But, for developing countries finance and availability of space to build such infrastructure might be a serious issue. Another approach is the use of sensors and cameras. But these have issues like limited coverage, expensive implementation and maintenance associated with them. Another approach is probe-vehicle data(Google MAP, Bing Map etc.). But these have issue like noisy data on the arterials. Another method suggested uses Vehicle to Vehicle Communication, but these require deployment of equipments in vehicles and have concerns like security and privacy associated with them.
As these cities are technologically advanced they harbor large population of social media users. Thus, we propose a method which uses resources which are readily available to these people. Social media data can be leveraged to solve traffic issue. As social media data is unstructured, we need to develop a interactive visual display which provides real insights into traffic problems. This data can not only be used by the common man but also respective traffic authorities to take actions to curb traffic. We analyse this social media data to not only get traffic-related information but also sentiment patterns. We can also perform statistical analysis on this data to compare traffic between two cities.

Traffic in the cities is divergent and is unpredictable. Road traffic congestion has always been a challenging task for the urban areas. Moreover, monitoring such heavy traffic using sensors or probe vehicles is also not feasible because of the high installation cost and certain environmental conditions. The traditional method for monitoring the traffic was by using sensors. Those sensing technology also consisted of intrusive and non-intrusive detectors. The non-intrusive sensors were highly skilled in sensing traffic speed , volume and classification as they included infrared, high sensitive, microwave radar, sound sensitive, ultrasonic and VIP sensors. Then a novel idea of monitoring the traffic using social media such as Twitter was brought up[8]. Twitter is a widely used Social media networking site. It is a rich source of

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

news from across the globe. This social media is a good means of spreading news at earliest possible. Social media is used as an awareness part. The long tail events can be detected because of sparsity. Also monitoring of fire events is done. The system is trained to check only for fire events. Text classifier is used to check to check for related event of interest and all the related events are grouped to form cluster. This information helps to alert the people living in the vicinity. A text classifier is used to identify tweets of interest including those accompanied with photos and monitoring system that can track multiple events at once.

II. LITERATURE SURVEY

Traffic in the cities is divergent and is unpredictable. Road traffic congestion has always been a challenging task for the urban areas. Moreover, monitoring such heavy traffic using sensors or probe vehicles is also not feasible because of the high installation cost and certain environmental conditions[7]. The traditional method for monitoring the traffic was by using sensors. Those sensing technology also consisted of intrusive and non-intrusive detectors. The non-intrusive sensors were highly skilled in sensing traffic speed , volume and classification as they included infrared, high sensitive, microwave radar, sound sensitive, ultrasonic and VIP sensors. Then a novel idea of monitoring the traffic using social media such as Twitter was brought up[8]. Twitter is a widely used Social media networking site .It is a rich source of news from across the globe. This social media is a good means of spreading news at earliest possible. Social media issued as an awareness part. The long tail events can be detected because of sparsity[9][10]. Also monitoring of fire events is done. The system is trained to check only for fire events. Text classifier is used to check to check for related event of interest and all the related events are grouped to form cluster. This information helps to alert the people living in the vicinity. A text classifier is used to identify tweets of interest including those accompanied with photos and monitoring system that can track multiple events at once.[11]

III. METHODS

A. Initial Data Search and Collection:

The data uploaded on social media by users can be accessed using the REST API(REPRESENTATIONAL STATE TRANSFER APPLICATION PROGRAM INTERFACE). Rest is probably the most simple way to use all the web services on data. It is fully-featured and relies on a stateless, client-server, cacheable communication protocol. It is not a standard which helps you to "roll your own" basically make use of it according to how you would like to use it with standard library features in languages like Perl, Python, Java or C#. The Client makes a HTTP request to the REST Web Services Server. The server then establishes a streaming connection and issues a request to the Twitter REST API. Twitter accepts the connection and streams Tweets as they occur. The HTTP Server pulls the data from the data store and renders them into view. This is the response the Client receives. The client can then see the tweets displayed on the website. The Twitter API also filters tweets with specified hashtags according to the need of the particular application. The REST uses HTTP for all the four CRUD operations of Create, Read, Update and Delete. OAuth(Open Standard for Authorization) is used for secure access of data. It issues tokens which are to be issued to the client by an authorization server on behalf of the resource owner. Thus, this data collected using the Twitter REST API is used for further for analysis. The Figure below shows two server processes, one server receives the streamed tweets, while the other brings the result and renders it into view.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

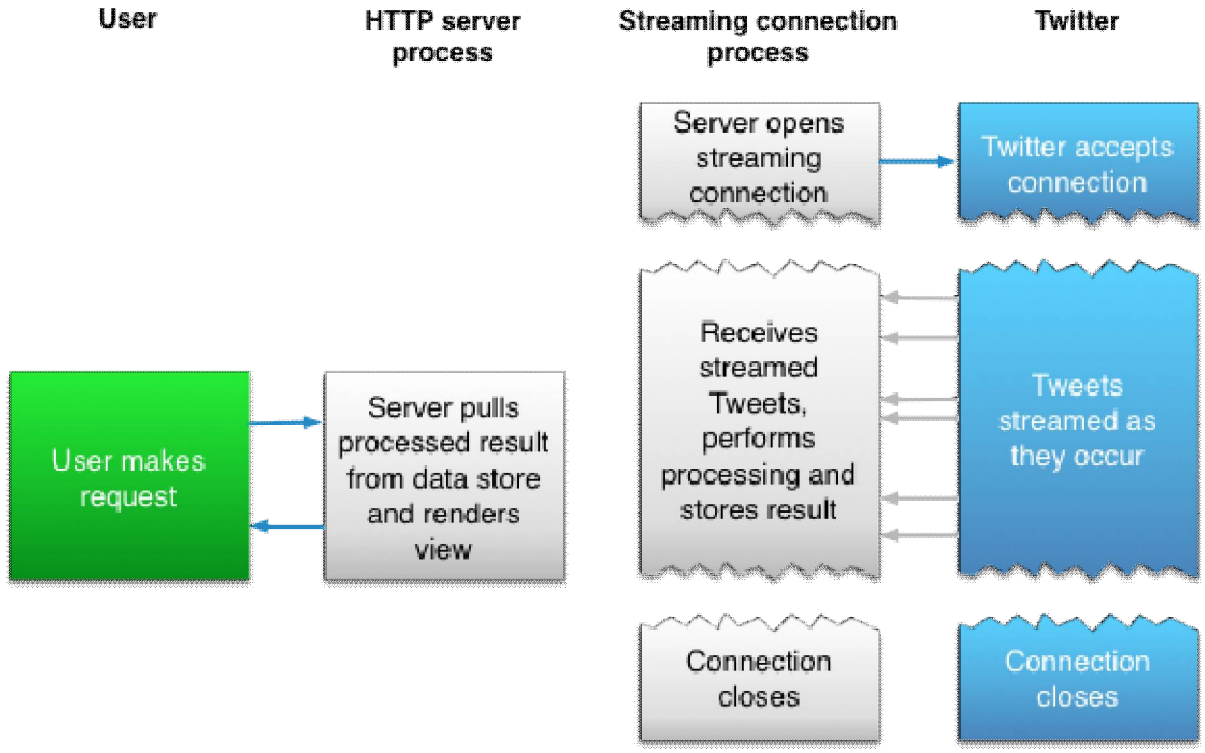


Figure: Two server processes and the flow between them.

B. Initial Text Mining and Preprocessing

Data Preprocessing is required as real world data is incomplete, noisy and inconsistent .i.e. it lacks attributes of interest, contains errors and discrepancies in names and codes. Data Preprocessing is an effective step for analysis of data. It removes unwanted, irrelevant and noisy data and helps in giving good quality data.

1. Data Extraction

A tweet has 140 characters max or less. The Twitter API helps us to perform data extraction by crawling through to retrieve only the relevant information from the data sources like database or web pages. Thus giving us only tweets of interest having particular hashtags. In this case, data extraction is used to obtain traffic related tweets. The contextual features like(time stamp, location) are seen during data extraction.

2. Data Cleaning

Data Cleaning is used to remove corrupt or inaccurate records from a record set or database. After performing data cleaning, all the records are consistent to the other records in the database. In this case we have performed data cleansing by removing tweets which have typographical mistakes and converting words to their base form using Stanford Language Processing Parser.

C. Data Analysis

The data that is obtained in the form of tweets contains information like the message in the form of character string, timestamp, location, user name, retweets, etc. This data obtained is stored in the database and subjected to preprocessing which yields only those attributes which are needed by the user or particular application. This preprocessed data is again stored into another table of the database and analysis is performed which consists of removal

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

of stop words, stemming and calculating word frequencies. Sentiment analysis is performed on this data to obtain the exact meaning of the tweet and what the user is trying to express through his tweets on the social networking sites.

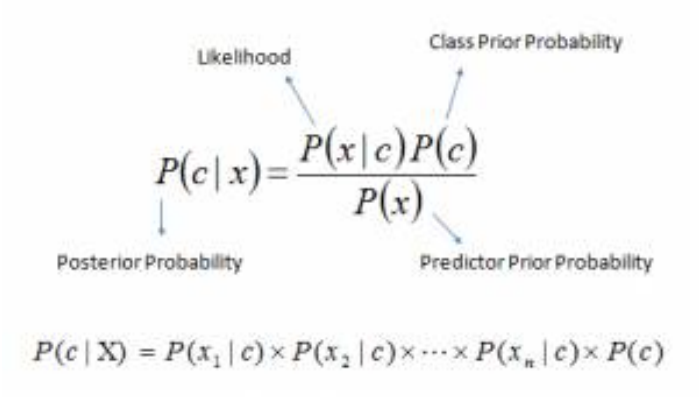
Sentiment Analysis

One of the major benefits of using a Social Listening tool is that you can quickly determine people's feelings related to any event, product. But Sentiment Analysis can be a subjective tool, and people with some opinion are often observed to identifying the positive, negative or neutral sentiment from a comment or review. Sentiment Analysis can be defined as to automatically extract or classify sentiments from mostly unstructured data or text using the combination of natural language processing (NLP) and the computational techniques. There are a few approaches which accelerates the processing of the analysis out of which Naïve Bayes is one of the efficient approach. It is very easy to construct as it does not require any complex iterative parameter estimation schemes. This specification clearly indicates that it may be readily applied to huge datasets. It is easy to interpret, so it even becomes easy to the users to understand who are untrained in classifier technology can understand. And finally, it often works surprisingly well. Though it cannot be called the best possible classifier, it is reliable in terms of it's performance and does quite well.

Naïve Bayes

Thus, despite the fact that Naive Bayes usually over estimates the probability of the selected class, it usually provides with the decision which further does not accurately predict the actual probabilities, the decision making is correct and thus the model is accurate.

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:



Above,

- $P(c|x)$ is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

In a text classification problem, we will use the words (or terms/tokens) of the document in order to classify it on the appropriate class. By using the "maximum a posteriori (MAP)" decision rule, we come up with the following classifier:

$$c_{map} = \arg\max_{c \in C} \left(P(c \mid d) \right) = \arg\max_{c \in C} \left(P(c) \prod_{1 \le k \le n_d} P(t_k \mid c) \right)$$

Where t_k are the tokens (terms/words) of the document, C is the set of classes that is used in the classification, $P(c \mid d)$ the conditional probability of class c given document d, $P(c)$ the prior probability of class c and $P(t_k \mid c)$ the conditional probability of token t_k given class c.

This means that for the classification of the document, estimation of the probability of each word of the document given a particular class (likelihood), multiplied by the probability of the particular class (prior). After calculating, the one with the highest probability is selected.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

D. Routes and Directions

The Google Maps API can be used for this purpose. We can integrate the features and services provided by the Google Maps API to obtain the routes and directions for travel along with the estimated time to travel. As the tweets do not provide the latitude and longitude of places particularly cities. We can maintain a database which dynamically updates as places are extracted from tweets. To check if the place mentioned in the tweets exists we can use Google Maps API for the same purpose. The place or city then obtained can obtain it's coordinates and store these into the database. This information can then be used to plot maps as one of the main requirements while plotting the map are the coordinates of a place. One can then enter the source and destination of the place they want to travel to and obtain the driving directions and estimated time travel.

IV. VISUALIZATION

The visualization of the Dashboard can be done using HTML and CSS or using the D3.js which is a Javascript Library used to manipulate data. The Dashboard displays the user tweets and the sentiment analysis of those tweets. It also displays the pie charts which gives statistics about a particular city's traffic on a particular day or during a particular month(according to how we design it). The maps can be used to get driving directions and estimate the time to travel from source to destination. The routes which are congested and shown on the dashboard can then be avoided and alternate route can be chosen from the map.

V. SYSTEM ARCHITECTURE

1. Data is collected from Twitter with the help of Twitter API v1.1.
2. Data is then pre-processed to extract useful data and clean the noisy data.
3. The data is then tokenized and stop words are moved, stemming is performed and word frequencies are calculated.
4. Sentiment Analysis is then performed on data to find out the mood of the users or sentiments of people of a particular city.
5. This information is then displayed on a Dashboard in the form of graphs and pie charts also showing the tweets of the user using a real-time Graphical User Interface Dashboard.

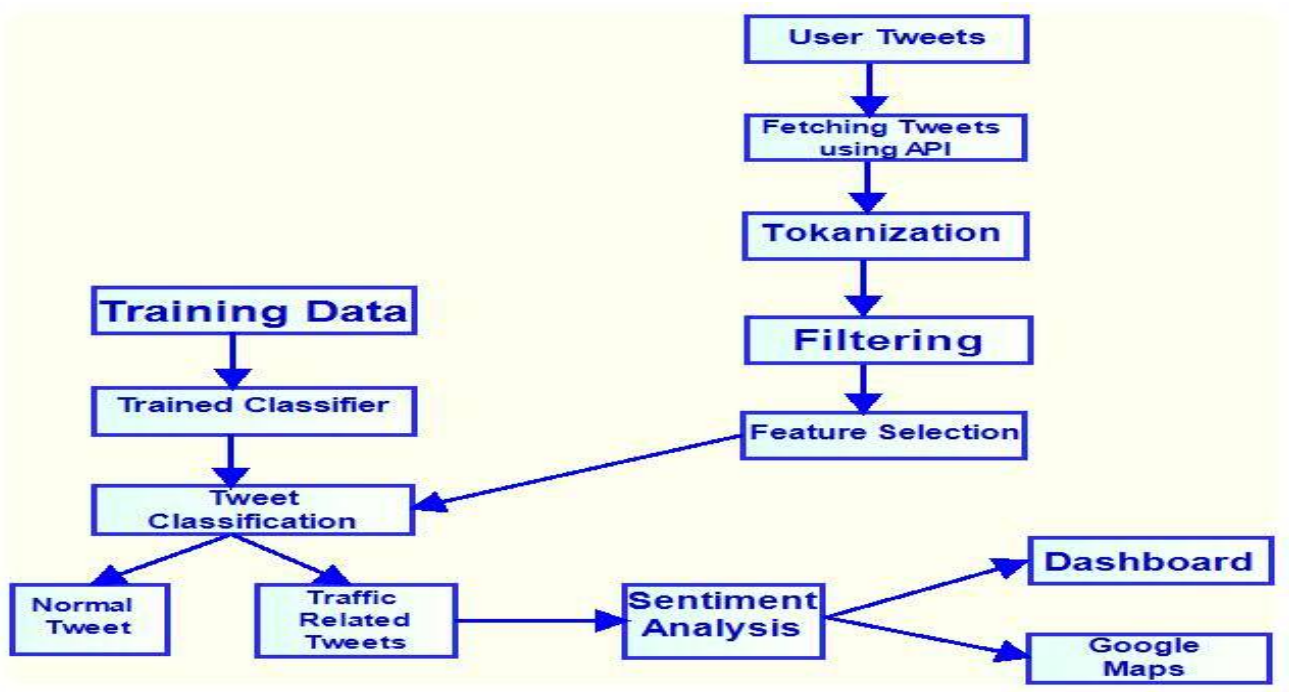


Figure: System Architecture.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

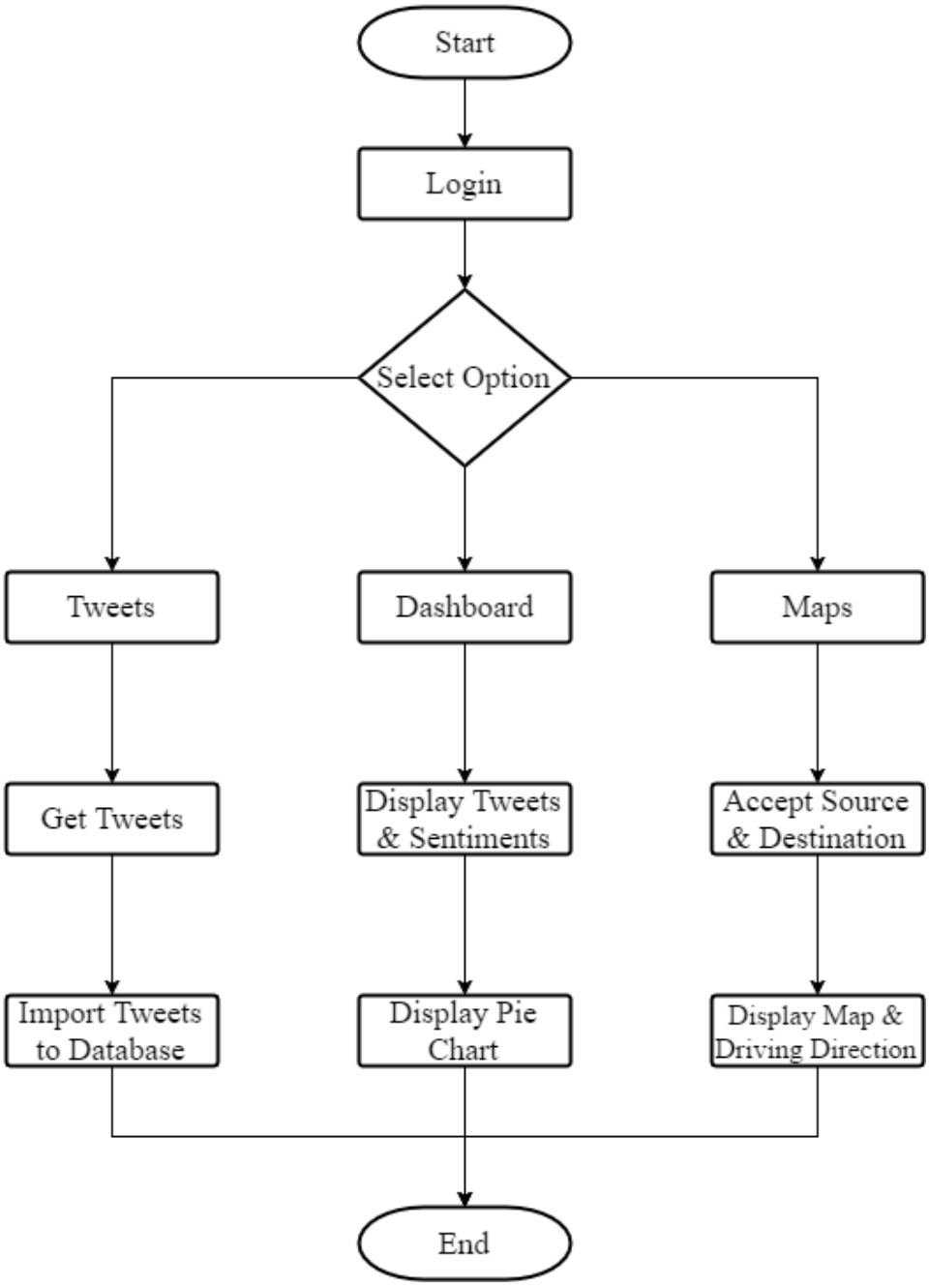


Figure: Flowchart.

VI. RESULT

We made a website where we kept a login page, one from where the users and the admin can login.

Next, we created another page where the tweets captured using the Twitter REST API are displayed. We used a grid view with columns which display information like location, date and time stamp of the post, user and the tweet message. We've enabled the paging so that large number of tweets can be displayed. There's also a Import Live Tweets button on the website. On clicking the option the tweets are imported into the database. We've also made sure that only

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

one copy of every tweets gets saved on the database and the duplicates are avoided. This we made sure by if the tweets are imported into the database a number of times, only the tweets imported on the recent most click are saved and the previous tweets of that day are discarded. At the end of the day all these tweets are saved into the database as final tweets for that day. Then, data extraction and cleaning is performed on these tweets and the new list of useful tweets are saved into another table of the database. The sentimental analysis is then performed on this useful list of tweets and each tweet with it's message and sentimental analysis is displayed. Also, the pie charts showing sentiment analysis of traffic on the particular day of a particular city is displayed. This is just one of the ways you can design your interactive Dashboard according to how you want it to look to increase aesthetics. Important is how you perform the analysis.

We also created another option on our Dashboard where you can use the maps for travelling directions. We used the Google Maps API functionalities and integrated it into our website. So now the user can not only see the route from it's route to destination and get the estimate of the travel time but also can avoid routes which are crowded or have negative sentiments associated with them on the Dashboard and can take an alternate route. Figure below shows the maps and the route from Pune to Mumbai and the estimate of travel time on the driving direction panel.

The efficiency of our system increases as time passes and more and more database or data is acquired as the whole process is based not only on the quantity of data but also the quality of data. As the system gets trained on more and more data and comes across more real-time results after their analysis the efficiency of the system increases and after a point of time it is estimated to perform more accurate than ever.

VII. CONCLUSION

This paper addresses the problem of traffic congestion by focusing on monitoring of the traffic using Twitter as the convenient social media. As the result of traffic monitoring is going to guide the public, the processing should be done with much more accuracy. So the related work mostly focus on large-scale event detection. Large amount of data related to any particular traffic related event leads to more precise idea about the event occurred, and thus helps with accurate results. In our work, we demonstrate how a language model and Naive Bayes classifiers can be combined to identify traffic congestion locations and display the statistics of sentiment analysis using a pie-chart. The system is able to retrieve and classify streams of tweets which naturally provides an alert signal to the users regarding the existence of traffic events.

VIII. FUTURE SCOPE

Data captured from the cameras can give us a more detailed view of the event occurred and where it has occurred with probable proof of the reason and give us the exact location as well as the cause. We can use this feature with our project to give a detailed information about an event, accident etc. The use of cameras fused with the project will not only manifest where the event has occurred, but can also give the reason as to why it occurred.

REFERENCES

1. Sentiment Classifier in PHP, https://github.com/JWHennessey/phpInsight.
2. Twitter API Wrapper for Twitter API v1.1 Calls, https://github.com/J7mbo/twitter-api-php.
3. FacebookGraphAPIExplorer,https://developers.facebook.com/tools/explorer.
4. Bo Pang, Lillian Lee, Opinion mining and sentiment analysis, 2008.
5. NaiveBayesClassifier,http://en.wikipedia.org/wiki/Naive Bayes classifier.
6. Google Maps JavaScript Library, https://github.com/hpneo/gmaps.
7. Yu, P. D. Prevedouros, Performance and Challenges in Utilizing Non Intrusive Sensors for Traffic Data Collection, Advances in Remote Sensing,2013, 2, 4550.
8. Jianshu Weng; Yuxia Yao; Erwin Leonardi; Francis Lee, HP Laboratories HPL201198 Event Detection in Twitter. In ICWSM, 2011.
9. P. Agarwal, R. Vaithiyanathan, S. Sharma, G. Shroff. Catching the Long-Tail: Extracting Local News Events from Twitter. In ICWSM, 2012.
10. Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Philip S. Yu, Hongjun Lu. Parameter free bursty events detection in text streams. In VLDB, pp. 181192, 2005.
11. Aiello, L. M., Petkos, G., Martin,. Sensing trending topics in Twitter , 2013.