UNIVERSITY OF NAIROBI

SCHOOL OF COMPUTING AND INFORMATICS

VISUALIZING NAIROBI TRAFFIC FROM SOCIAL MEDIA DATA

BY

MUTUA JACKSON MULINGE

P58/75731/2012

SUPERVISOR

PROF. PETER WAIGANJO WAGACHA

A project submitted in partial fulfillment for the award of the Degree of a Master of Science in Computer Science at the University of Nairobi.

July, 2016

DECLARATION

This project is my original work and to the best of my knowledge this research work has not been submitted for any other award in any University.

Mutua Jackson Mulinge: _____ Date: _____
(P58/75731/2012)

This project report has been submitted in partial fulfillment of the requirement of the Master of Science Degree in Computer Science of the University of Nairobi with my approval as the University supervisor

Prof. Peter Waiganjo Wagacha: _____ Date: _____

School of Computing and Informatics

ACKNOWLEDGEMENT

To God, Thank you for this wonderful opportunity, I am forever grateful. To family and friends, thank you for your support all through. Allan Kibet, many thanks. Prof. Wagacha: We did it, from scratch to here, your guidance, ideas, criticism has made everything a success. Dr. Orwa, the push and encouragement has made it happen. The entire panel, Digital Matatu Team, nrb.city and *Ma3Route* crew I am and will always be thankful to you.

ABSTRACT

Traffic congestion is one of the greatest challenges faced by developing cities in the world. Nairobi, Kenya is not exceptional. Traffic data collection methods such as cameras, radars are too expensive to install and maintain or are unavailable. Citizens have turned to social networks and application such as ma3route to share traffic information. In this paper, we mine traffic information from tweets send to @ma3route handle using the array explode method. Most tweets received on ma3 route correspond to a particular road status. The tweet is parsed and processed. The mentioned location is then paired with the traffic condition information and is visualized using a map application such as Google maps. All this processed tweet data is stored in a database. The traffic status for a specific location for a specified period can be displayed upon query execution. We are able to analyze the traffic condition of a specific location for a 24 hour period and visualize it, enabling us to see traffic patterns such as peak periods. We verified our processed data against on-the-ground checks and crosschecked against other data sources such as @RoadAlertKE and @kenyatraffic and found that the data matched. We conclude that social media data provides an alternative, cheap and real time source of traffic data.

.

Table of Contents

TABLE OF FIGURES

LIST OF ABBREVIATIONS

API - Application Programming Interface

ATIS – Advanced Traveler Information System

CCTV – Closed Circuit Television

CSV- Comma Separated Values

CSS – Cascading style sheets

FCD – Floating Car Data

GPS – Global positioning system

GSM – Global System for Mobile communication

ITS – Intelligent Transport System

JSON - Java script Object Notation

HTML – Hyper Text Mark-up Language

NER- Name Entity Recognition

OOV – Out Of Vocabulary

POS – Part Of Speech

RAID – Risk, Assumption, Issues and Dependencies

SQL – Structured Query Language

SVM – Support Vector Machine

TMC – Traffic management centre

CHAPTER 1: INTRODUCTION

1.1 Background

A country's economic, commercial and social progress can be quantified using transport as a measurement feature. Traffic congestion stands out as one of the greatest challenges faced by cities in the world in developing countries. Valuable time wastage, increased carbon emissions, loss in monetary terms and reduced productivity in work force are negative impacts of traffic congestion. Population increase, lack of mass transport systems and poor road infrastructure contribute to increased road congestion. Losses of up to 5.5 billion hours and 2.9 billion gallons of fuel are wasted in traffic snarl-ups which translate to a total $121 billion yearly (Lecue et al, 2014).

Usefulness of public transport can be improved through provision of reliable transit information. Collection of data on urban transport study involves placement of sensors, roadside surveys, induction loops and use of radars. These methods are effective for the purpose they are designed for however harbor several shortcomings which include expensive installation and constant maintenance, labor intensive, limited coverage area and collection of specific data type (Carvalloh, 2010).

Micro blogging has changed the way in which citizens share information. It has emerged as a rich source of multiple types of data and presented a new interaction platform (Elsafoury, 2013).Twitter and Facebook are some of the social sites where users post and share information about issues affecting them. Instances in which twitter has been used to relay accurate real time information are on the increase (Carvalloh, 2010). Knowledge of current traffic status can help in travel decision making.

Nairobi, Kenya is no exception to traffic snarl-ups experienced on a daily basis. Road users share this information on social media such as Twitter and Facebook. Several twitter handles on which users post traffic information include @ma3route, @RoadAlertsKE @myroadtraffic and @kenyatraffic.

Application developers have come up with transport related apps like *Ma3route* and *Kijicho-app* which enable road users to post traffic status on their current roads. *Ma3route* application enables users to share traffic information (Klopp et al, 2015). GPS traces from devices that can access the internet can be used to relay real time traffic information (Boulos, 2011). With such advances on adoption of social media use and data generation, people have been termed as human sensors. Social media has emerged as a source of real time data.

1.2 Problem statement

Traffic congestion remains one of the biggest challenges experienced by developing countries. In order to understand and develop transport systems, accurate and reliable data is essential. Collection of transport data is key in designing useful and reliable transit systems. Nairobi lacks a real time traffic monitoring system therefore relying on *Ma3route* updates to share data (Santani et al, 2015). Methods employed in collecting data determine its usefulness and accuracy. Use of road side surveys, induction loops, cameras and radars in data collection are expensive, collect a specified data type, inaccessibility of data after collection by the designated agencies and sometimes the data is not real time (Carvalloh, 2010). With the increased demand for data, alternative and more affordable data collection methods are definitely desired. Social media and GPS enabled devices have in the recent past enabled human beings become smart sensors. GPS enabled mobile phones can be used as data collection tools to provide important transport information (Klopp et al, 2015). Data from these sources have Geo-tagging feature which provide the location of the user or post if a user has enabled it (Elsafoury, 2013). Twitter presents a cheap, easily accessible resource to mine traffic information. Ability of social networks to provide real time data is an added advantage over traditional methods of data collection. If people get real time information on the traffic conditions on the roads, they are able to make informed decisions including alternative routes or changing their schedules to avoid been held up in traffic for long hours.

1.3 Objectives of the study

1) To collect and mine social media data for traffic information.
2) To establish existing traffic visualization systems from social media data.
3) To develop a prototype to visualize Nairobi traffic on maps and use it to determine traffic patterns at different times of the day based on data updates.

4) Identify congestion areas or points of interest based from the visualization.

1.4 Research questions

1) Which tools are used in collecting and mining information on social media?
2) Which traffic systems exist for visualizing traffic?
3) How will a prototype to visualize Nairobi traffic be developed and use it to establish traffic patterns in Nairobi based on social media updates?
4) How can congestion points in Nairobi be identified?

1.5 Justification of the study

Nairobi city has been described as the fourth most painful city to travel in the world based on traffic congestion among other factors (IBM report, 2011). Availability of real time traffic information is crucial to avoid congested areas (Elsafoury, 2013). Lack of proper planning during development phases, increased number of road users and lack of mass transport systems are some of the key causes of traffic snarl-ups. Several studies have been conducted in developing Nairobi transport model with road side surveys been the most used. Other methods like cameras and radars are too expensive to install and maintain. Availability of affordable smart phones has enabled people become human sensors. With the increased use of social media, a lot of data is being posted. It presents a cheaper and real time source of traffic information. Users express how they feel through their status. Results of the study can be applied to other towns.

1.6 Scope of the study

The study covers Nairobi area roads based on traffic updates by users. Its main source of data is *Ma3Route* application and Digital Matatu project data. *Ma3route* data is received from various sources such as sms, direct posting on the mobile and web application and @Ma3Route twitter handle.

1.7 Limitation of the study

A lot of data is posted on social media websites. The study relies on information posted by the users. The validity of this data in terms of time and traffic status presents a challenge since it's based on the user's judgment of the traffic situation. Most of the tweets are grammatically

incorrect and use slang language. There is no standard sentence structure for posting data thus presenting a challenge during processing. The study is limited to Nairobi area.

1.8 Data limitations

Most of the users have not activated Geo locator on their devices hence the tweets posted *@ma3Route* do not have the geo location for fear of been tracked. To obtain a geo location, we had to work backwards from a mentioned or described location to obtain an approximate geo location. Though there are a large number of tweets processed, some of the tweets are not related to traffic but rather the trending topics and are not useful in our work.

CHAPTER 2: LITERATURE REVIEW

2.1 Sources of traffic data

2.1.1 Social media

Social media has changed the way people around the world communicate. Twitter, since its launch in 2006 has an estimated 44.5 million users with a monthly growth of 1382% users. An estimated 2.55 billion people will be social media users by 2018 around the world (Statista, 2016).

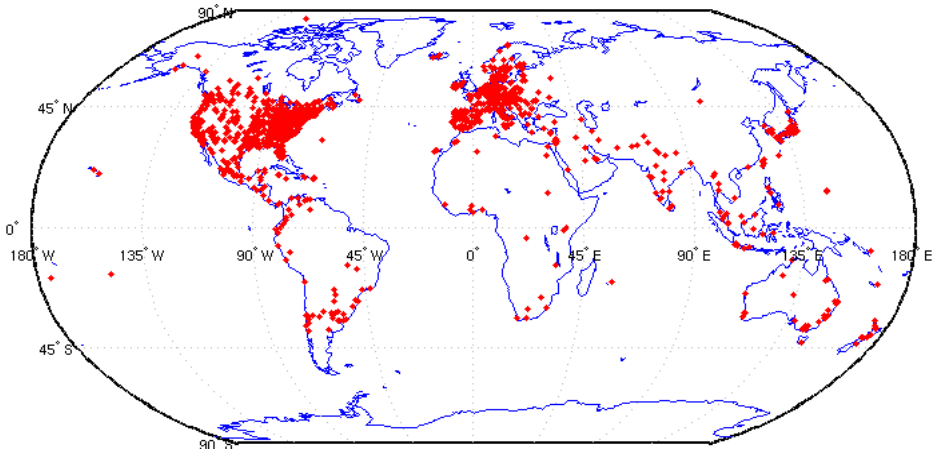


Figure 1: Twitter users distribution

(Source: Sakaki et al, 2010)

Tostes (2014) observed that Foursquare and Instagram experienced enormous growth in 2014 with Foursquare registering 45 million users, Instagram 200 million users and Waze 50 million users respectively. Generally these applications are population dependent for information hence the term crowd sourcing hence the more user they have, the higher possibility of more data being posted (Boulus, 2011). Transport agencies own twitter accounts such as @TfLTrafficNews for London traffic updates (Elsafoury, 2013). Data mining tools can be used to harvest traffic information posted on twitter.

2.1.2 Road side sensors

Traffic studies around the world involve the placement of a network of road side sensors such as inductive loops, cameras and radars in collecting information from a particular point (Tao et al,

2012). Several challenges are associated with this data collection method. The stationery and fixed nature of this method means only data for that particular point is collected therefore cannot be used to predict the behaviour of the entire road network. High cost of installation also limits large scale deployment and also dedication to a particular type of data limits their capability in establishing general traffic performance (Carvalloh, 2010, Tao et al, 2012).The data collected is not real time and requires post processing to increase its usability.

Road side surveys were traditionally used in traffic data collection but with increased technology adoption, it's used sparingly and on specific occasions. It is very tedious, time consuming and requires a lot of man power. A lot of processing is required on data collected and accuracy is purely based on the surveyor's observation.

2.1.3 Automatic vehicle location

Automatic vehicle location (AVL) involves determining the geo-location of a vehicle and relaying that information to a database (Portillo, 2008).Data generated as the vehicle moves around is stored in centralized database. Using the geo-locations and time when they are transmitted to the server, traffic conditions can be generated from that data. Tao et al (2012) describes the expensive hardware installation and maintenance of the equipment for communication between the vehicle and database centre as a limitation in large scale adoption of AVL.

Cost of purchase, installation monitoring and maintenance of equipments to support these methods limits them in terms of their application as traffic data sources. Also the input required in processing the data is enormous. Data is most useful when it's most recent and therefore some of these methods do not relay data in real time reducing their application range.

2.1.4 Crowd sourcing information

Crowd sourcing is viewed as an effective method of generating masses of real time data. Howe (2006) defined crowd sourcing as the act of distributing a job that was originally preformed by a single person to a disjoint mass of people often termed as an open call. World Wide Web has been used in developing and supporting crowd sourced applications (Doan et al, 2011).

Crowd sourcing involves the use of collective wisdom from a crowd to provide a solution to a problem affecting that crowd (Misra et al, 2014). It encourages public participation on policy

making. Given that it is the crowd who are the end beneficiaries of crowd sourcing, a lot of data is posted thus making it a data repository. Google's *Waze*[1] is an example of a crowd sourced data application where users share traffic updates which available in real time to other users thus saving other people from getting stuck in traffic.[i]

However several drawbacks are experienced with this data including the lack of implementation of the information shared by the public. Most of the data posted on these platforms is poorly structured therefore aggregating it is quite difficult. Analyzing data obtained from crowd sourcing is a big challenge also due to scalability of the underlying algorithms and also due to structural complexity. Domain specific systems also lockout some of the participants.

2.1.5 Mobile phones as a data collection tool

Advanced technology in mobile world has facilitated the use of mobile phones in data collection (Tao et al, 2011).Availability of affordable smart mobile phones that have both GPS and internet connectivity capabilities which can send their current Geo location has increased the data collection tools. Mobile phones have automated data collection methods. The results are less paper work and processing time since right information can be captured at collection point.

Muthiah's (2011) study used the Mhealthsurvey to transmit health records for patients in India. He acknowledges that real time data transmission and reduced work load are some of the benefits of using mobile phones. A wide range of applications available on Google play store for android phones and Ovi store for Nokia phones use GPS on those phones to transmit data. The *Mytracks* [2]application was used during the Digital Matatu data collection exercise and offered more capabilities compared to the hand held Garmin devices used for the same course (Klopp et al, 2015). *TransitWand*[3] application records the travel time between stops, average speed and total time during the entire journey therefore using the data generated, traffic congestion can be established.

[1] Waze Available at: https://www.**waze**.com/

[2] MyTracks available at: *https://en.wikipedia.org/wiki/**MyTracks***

[3] TransitWand available at : ***transitwand**.com/*

Most of social media users access their accounts via mobile phones. Therefore the increased data transmission can be attributed to GPS and internet enabled phones.

2.2 Related social media studies in real live events.

Sakaki et al (2010) used twitter messages to develop an earth quake reporting system for Japan. Support vector machine (SVM) was used to filter earthquake related tweets whether as positive or negative. Morphology analysis using Mecab was used to separate words in a sentence to a group of words for tweets sent using Japanese language. Keywords such as "Earth quake" or "shake" are used to identify related tweets. The SVM was used to classify the tweets as positive or negative. In this case, positive meant tweets that were connected to the Japan's earthquake. A probalistic spatiotemporal model was developed for finding a centre and trajectory of the target event. Sakaki (2010) and the team used the Kalman filters and particle filters for estimating location. The system was used to send emails to people alerting them of earthquakes based on twitter information.

Dork et al (2010) approach presented the visual back channel concept which helped in following real world events. It comprised of topic streams, people spiral and image cloud providing several modes to visualize large scale events. A lot of text analysis is usually done to incoming raw tweets. The text analysis procedure includes collecting re- tweets and photos, removing noise from data, stemming words with same meaning and in the end associating extracts to the original tweet. The visual backchannel implements its sub components finally presenting a clear picture that is easily understandable on the trending topic.

Asur and Hurberman (2010) used twitter in predicting real world outcomes. They predicted the revenues from box-office movies using twitter chatter. The initial focus was on how much publicity was created for each movie and its trend as time progressed. Next step in their study was to establish how sentiments are generated, how the viewers are affected by good and bad opinions about a particular movie. Sentiment analysis was used together with text classifiers to differentiate positive tweets from negative ones. The results from the linear regression model developed showed that the system's accuracy surpassed that of the Hollywood stock exchange in terms of revenue prediction.

The above studies show that twitter is a resourceful real time source of data. Using several methods and algorithms it is possible to develop systems that collect, transform and relay information in an easy, understandable and interpretable way to users and also help policy makers in decision making.

2.3 Traffic visualization systems

2.3.1 Floating car data system

Kiu et al (2009) adopted the use of Floating Car Data (FCD) as a source of data in large scale traffic monitoring. Cars are fitted with GSM and GPS systems that provide location, speed and direction and are used as data collectors thus overcoming the disadvantages of using radars, loops and cameras in road data collection such as delays in data collection, coverage area, and cost incurred (Hubber, 1999. Kiu, 2009). Data is collected for the entire road network as the vehicle moves along therefore decisions on overall roads can be made based on it rather than segments.

When the FCD feeds are received they are mapped to road network and queries that relate to that road network are updated. Taxi data from the city of Beijing was used due to their extensive movements around the city and produced animated traffic visualization. FCDs challenge is presented when excess feeds are collected that overwhelm the systems processing capability thus resulting to increased delay processing time and reduced real time monitoring of traffic status. Such situations forces the system to drop single queries with high cost and adopt many queries with minimum cost thus maximizing on the overall number of queries provided in a process called load shedding.

2.3.2 The web marsh-up traffic visualization system

Piccozi et al (2013) used the city of Oulu traffic data set to present a mash up visualization using various data mining, web tools and visualization techniques. Key to the system is a clear understanding of the traffic states at different times of the day, week and on specific occasions. The data is based on the traffic counters thus number of vehicles in a given time frame. The developed system is made up of both a data miner and data web mash up front.

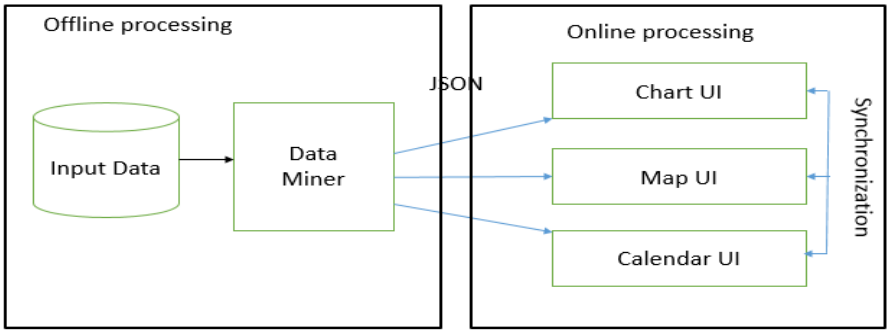


Figure 2: Web Marsh up system architecture

(Source: Piccozi et al, 2013)

Each intersection is denoted by a unique ID which is used to group data related to it and also condition at a given time. The generated data is from raw data is fed in JSON format to the front end. The user interface displays a city map showing circles at all the intersections each colored differently depending on the traffic intensity at that particular time, a chart and a calendar all showing details in regards to the selected time period.

2.3.3 4 Dimensional interactive visualization system

Pack et al (2005) incorporated OpenGL and several techniques to develop an interactive 4 – D interactive system using GIS, available transport data and in collaboration with real –time data management centre to analyze and visualize traffic data. The system is available to both public travelers and decision makers therefore increasing it usage. Pack (2005) argues that standard GIS systems can only visualize in 2 – D while 3 – D systems are slow in visualization thus the effect of real time is not achieved. Area of coverage included Washington D.C, Virginia and all of state of Maryland.

Packs system is composed of three stages; the terrain modeling, road feature mapping and dynamic data visualization. The images based on the visualizations cover the entire area thus acting as virtual cameras in all areas.

All the above systems require high capital cost and technology knowledge to ensure their implementation. Some cities and states have established dedicated traffic data centers thus ensuring data is reliable and timely. Advanced technology infrastructure is required to support these systems which are not locally available.

2.4 Visualizing traffic using social media data

Advanced traveler information system

Kumar (2005) defined an intelligent transport system (ITS) as one which relies on emerging computer, communication and information technologies in availing crucial information to system users in regards to traffic, route obstructions and alternatives and also cautionary messages. An advanced information traveler system (ATIS) is a system that relays relevant traffic information to commuters. Kosala (2012) developed an ATIS system which is a subset of an ITS that extracted and represented traffic information from different twitter pages on to a map to help the public gain a better understanding of the traffic situation in the city of Jakarta.

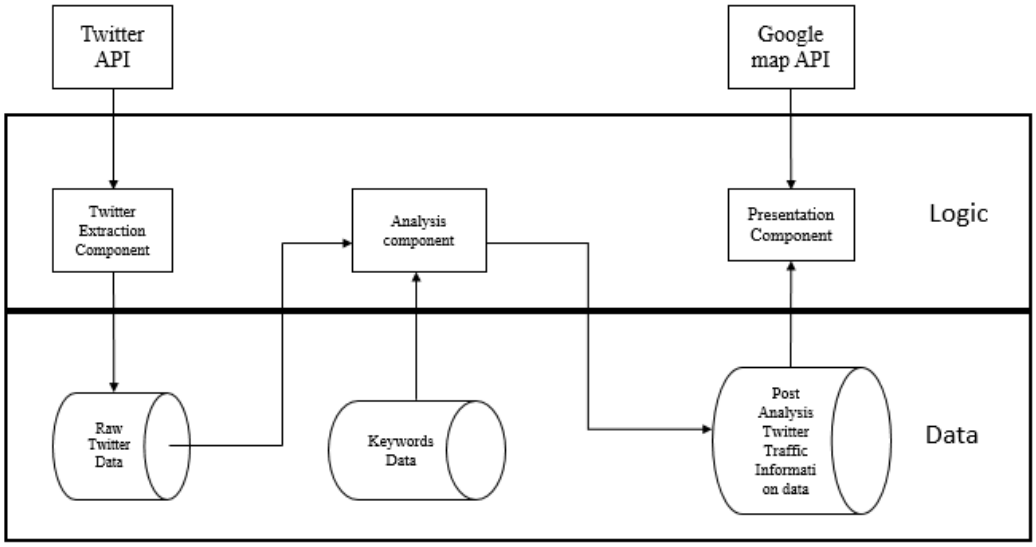


Figure 3: ATIS system

(Source: Kosala, 2012)

The system is composed of the following

i. Twitter API which is used by the system to query tweets falling under a particular criteria
ii. Twitter extraction component which holds results from twitter API query, parses the information in JSON format to the desired format and stores the data in SQL database.
iii. Raw Twitter data is a database where tweets returned from twitter search API are stored without modification.

iv. Analysis component loaded raw data from twitter and key words to do multiple layer analysis.
v. Key words data base stores the key words used to search tweets. The database contains multiple tables which include location table, abbreviation table among others.
vi. A post analysis data base which stores information derived from raw twitter database
vii. The presentation component which query's the post analysis data. It determines the confidence level of information.
viii. A Google map API presents a map that allows the processed data to be availed to the user.

The analysis component performs three types of analysis in parallel. Initially, it loads tweets analyses for keywords and if there are any changes the database is updated. Secondly, an abbreviation analysis is performed since in most social media it is a common practice to abbreviate. The failed tweets are loaded and a search for keywords is done, if a new word is found, the database is updated. Finally the system looks for location and traffic condition pairs them for each particular tweet.

A system confidence level defined as the level of confidence the system has with particular traffic information. With the information being real time, the confidence level reduces as time progresses from the time a message was sent.

Confidence Level = ((Time Constant Time Difference) / Time Constant) * 100%

The system represented an overall 76.13% accuracy level which is an analysis of comparing live CCTV footage and what the system generated.

Tostes (2014) used data known as check-in from Instagram and Foursquare to analyze traffic in the city of Manhattan. The check-in data from these applications was used to probe and understand traffic conditions.

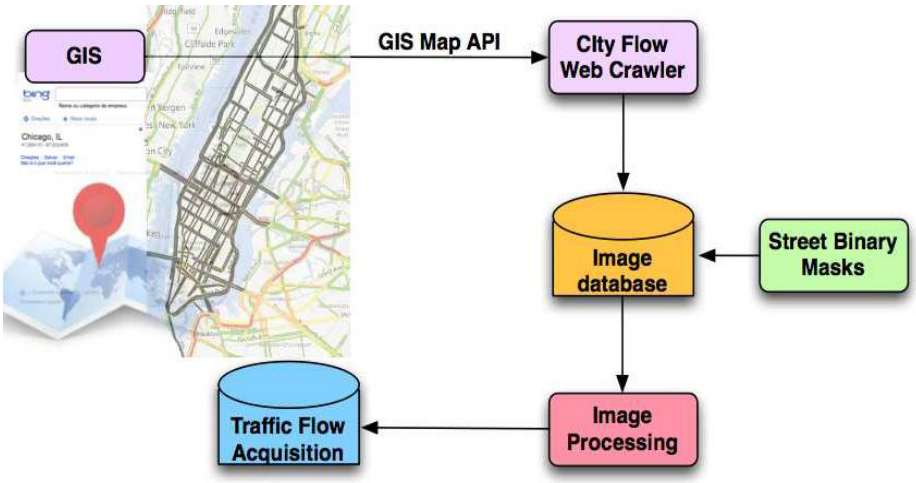


Figure 4: GIS map web crawler

(Source: Tostes, 2014).

City travelers share their locations through messages and photos. The check in are usually data shared from both Instagram and foursquare. Data for these applications is mined directly from twitter and only the tweets with latitude and longitudes are considered.

An API map service was used to develop a city flow web crawler and designed a bash script to collect the traffic flow image from the selected area. Virtualization was used in printing the screen which was processed using image processing software that extracted each road and finally saving the road intensity as a percentage of pixels with the highest represented by red and green by lowest. Results and analysis provided by the system showed that data mined from social sites and real traffic conditions were correlated. Social sensors distribution was equal to the traffic distribution with a variance **d** which was calculated to be thirty six minutes.

Carvalloh (2010) used support vector machine (SVM) identifying tweets that contained traffic information. He relied on messages sent by robots users due to the strict format to develop training set for the system thus overcoming the challenges of human sent messages which contain spelling mistakes, non standard punctuations and emoticons. The boot strapping approach was used. The initial classifier was used to identify positive and negative examples. Examples related to traffic are categorized as positive else negative. The positive training examples are added to the training set for a better classifier.

Elsafoury (2013) used the T-Pos algorithm developed by Ritter et al (2011) in classifying traffic tweets. The T-Pos tagger used the Stanford POS tagger with enhancements to overcome the problem of OOV (Out of vocabulary) by allowing use of #hashtags and @username symbols. The T-POS tagger generated a set of tagged tweets which are compared to a predefined dictionary for traffic and status. Three types of dictionary for normal, crowded and jammed for both traffic and status were used. In geo locating tweets, the T-POS tagger used customized dictionaries for extracting locations. The process involves extracting all nouns in a sentence which are then compared to a local dictionary; if they match a street name then it sends to the Google Geo-code API to get the geo coordinates of the point. The results are lines on a map for those streets with different color coding representing the traffic status of those areas.

Endarnoto's (2011) android application used tweets from traffic management centre (TMC) handle to visualize the city of Jakarta, Indonesia traffic. NLP was used to extract information from the tweets.

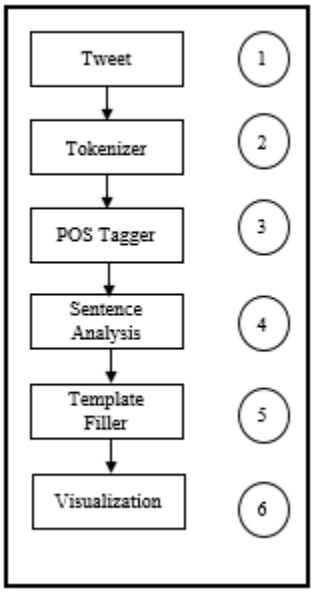


Figure 5: Information work flow diagram

Source: Endarnoto, (2011)

The tweets from TMC handle were saved in a database, then fetched and broken in to small parts known as tokens. POS is used in tagging the tokens into groups such as nouns, adjectives, noun phrases according to some predefined rules. The key elements sought for are Time, Origin, Destination and Traffic condition. The information is displayed on a map view by an android app. Displayed data is data extracted from the TMC twitter handle and processed. Three colors are used to represent the traffic conditions depending on congestion levels.

2.5 Text processing for social media data

A lot of data is posted on social media. To make value from it, text processing is performed to establish its usefulness in regards to a particular area of interest. Methods used in text processing of social media data include:

2.5.1 Name Entity Recognition

Collobert et al (2011) shows that Name entity recognition (NER) main objective is to categorize words in a text into given groups. Name entity recognises and classifies names entity in texts (Budi, 2012). NER is made up of several tasks with the major ones outlined below

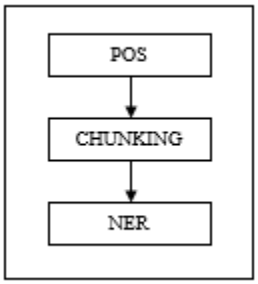


Figure 6: NER pipeline

(Source: Elsafoury, 2013)

NER tasks Include:

a) In POS, each word is assigned either as a noun, adverb or verb.
b) Chunking : It labels parts of a sentence syntactically into Noun phrases or Verb phrases

Two methods can be used:

- Rule based approach which classifies segments of a sentence according to some rules
- Supervised machine learning – a training set of labeled data is used to classify the text segments

c) NER: Actual classification takes place here. The classification process involves two approaches:

Rule based approach which uses written rules to classify entities. This approach uses two methods

- Look up – where only entities found in the Gazetteer are recognized.
- Language based – entity recognition is based on language rules.

Machine learning approach is more focused on classifying entities than recognizing them.

Hybird approach which combines both the rule based and machine learning thus resulting to increased accuracy in entity recognition.

2.5.2 Support Vector Machine

Support vector Machine algorithm is able to train a system to automatically categorize items according to a training set (Guduru, 2006). It is categorized as a supervised model used in classification and regression that uses (attribute, value) pair containing the (predictor, target) pair to establish the predictor and value relation (Elsafoury, 2013). It has two approaches

Linear classifier and Non linear classifier

In linear classifier the data is linearly separable. A hyper plane is used to separate one set from the other and similar data falls on one side of the hyper plane.

Non linear classifier the data in the training set is not linearly separable thus the output is of a polynomial shape.

2.5.3 Parts Of Speech

POS labels each word with a tag showing its syntactic role in the text i.e. if the word is a Noun, verb or adverb. POS tagger software available at http://nlp.stanford.edu/software/tagger.shtml#Download

performs this task. According to Elsafoury (2013) POS tagger has three major steps as shown by Robinson (2009). They include:

Tokenization which involves splitting the texts into small chunks known as tokens

Ambiguity look-up: This step places the most suitable tag to the unknown words

Disambiguation: Probability of a word being a noun or verb information is used to tag it.

SVM is a classifier rather than an extracting tool. Sakaki et al (2010) and Carvalloh (2010) experiments both used SVM in identifying whether a tweet is relevant to their researches or not. The name entity relationship classifies (NER) words in a sentence into categories (Collobert et al, 2011). Despite its high accuracy in its recognizing entities, Elsafoury (2013) noted the tool cannot be trained to identify new entities like traffic status.

The array explosion method will be used in this study due to the unstructured nature of the data being processed. This method breaks the tweet in to chunks similar to tokenization in parts of speech. The explode function returns an array of strings which are then compared to the search key.

2.6 Conceptual design

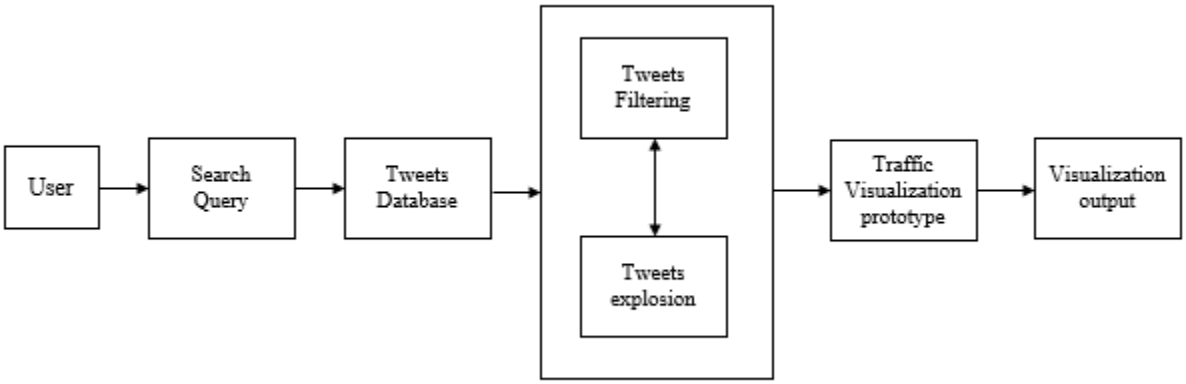


Figure 7: Conceptual design

(Source: Research)

The user issues a query to the system. The query can either be for location or accident statistics. Time specified by the user is used to set search boundaries. All tweets within the specified time are exploded and compared to the search key defined by the user. If a match is found, the mode traffic status of the given location is tagged to a pin posted on Google maps. If the query is about accidents, the number of times a vehicle type is mentioned within the specified time period is represented as a percentage of the total number of accidents for the bound period.

CHAPTER 3: METHODOLOGY

3.1 System development methodology

System development methodology is the set of proposed method for developing a complete system or its components while relying on logic and given philosophy (Yaghini, 2009,. Avison, 2006). Several methodologies are available each with a defined set of tools, standards and models but the choice of each is depended on the system requirements. RAID development methodology is used for developing this prototype.

Prototyping is the development of a replica of the real system whether as full or part of it in attempts to establish its feasibility and design characteristics (Gordon, 1995, Beaudouin, 2003). They are useful in envisioning the final system and mostly evolve into final product. Rapid prototyping involves two methodologies the throw away and evolutionary. In throw, the prototype is discarded after use while in evolutionary, the prototype forms the basis of the final product. Use of throw away prototypes is not economically feasible especially for minute projects (Gordon, 1995).

Evolutionary methodology was used for developing this prototype. The initial prototype undergoes changes during the development phases to become the final product. Such a methodology saves on time and other resources.

3.2 Prototype development

The prototype is developed using Boostrap framework. Often regarded as the fastest growing framework it combines CSS, HTML and JavaScript extensions used to develop dynamic web applications. It is preferred to other options because it is open source and free, enables the user to choose which components to install and can be customized to fit individual demands. It's available for download at http://getbootstrap.com/. Documentation and online support is available for users. Bootstrap is used in rapid development of user interface.

The array explode method is applied in extracting data from tweets. It is similar to tokenization used in POS approach where a string is subdivided into individual words. Its basic outline is:

array **explode** (string `$delimiter` , string `$string` [, int `$limit` = PHP_INT_MAX])

An array of strings which are substrings of the main string are returned. The string delimiter sets the split boundary. When a positive limit is set, the returned array elements are equal to the limit with the last element being the rest of the string if the string has more elements than the limit. Else when a negative limit is set, all elements are returned except the last element and when the limit is zero it is treated as one (Source: PHP manual)

3.3 Data source

Data used for this project is sourced from *Ma3Route* which is a traffic oriented application relying on crowd sourcing. *Ma3Route* has several platforms for collecting data which include its twitter handle @Ma3Route, a mobile phone and web application. Users post information about traffic conditions on their current location. Its twitter handle has approximately more than one hundred and twenty thousand followers. Users post a lot of data in regards to different topics. Data processing algorithms are used to filter traffic related data and the processed data is relayed to users using the same channels such that all the data appearing on these platforms is similar.

Digital Matatu data is freely available on GTFS exchange website i.e http://www.gtfs-data-exchange.com/agency/university-of-nairobi-c4dlab/ It contains all the stops in Nairobi together with Geo location for each. It serves as the local dictionary for this project.

Nrb.city is another source for data used in this project. It contains road traffic tweets for Nairobi.

3.4 Challenges in the algorithm

The sample was made of fifteen thousand tweets collected in the month of May 2015.A tweet is associated with a particular status i.e. Bumper to Bumper, clear, moderate or general information when users want to pass message to the public. The messages are not structured in a particular method therefore extracting relevant information is difficult. Several users can refer to the same location but in different formats. Example:

"Msa rd was bliss today. Mlolongo to nyayo in 30min"

"Mombasa road smooth today took me 15min to nyayo"

Both tweets refer to a similar route using different abbreviations and sentence structure. The local database must be expanded to accommodate all variations of sentence structure.

When users mention a road or route, the assumption that the whole route is congested or clear may be incorrect as only a part or particular parts of the route may be congested.

When words that are not location names are classified as such especially when a tweet is sent in slang language ‘Sheng’, this results in the system searching for all the assumed locations in the database and also Google maps API therefore increasing execution time.

3.5 Geo locating tweets

After array explosion, all words in a tweet are compared against a local dictionary. The local dictionary contains data from Digital Matatu project in which all locations are associated with their corresponding Geo coordinates. All words are viewed as locations which are compared to the local dictionary and if they match that particular Geo location is picked and a marker is placed on a Google maps else its sent to Google geo code API to get the latitude and longitude for that particular location. If a user has enabled location sharing on their device, the post usually contains the latitude and the longitude thus locating the source is simplified.

3.6 Traffic Status evaluation mode

Each traffic tweet is associated with a particular road traffic status. The study is limited to ‘Bumper to Bumper’ ‘Moderate’ or ‘Clear’. The tweets within the user defined time could have different status. The status of these tweets is computed and the most common i.e the mode status is selected as the overall status for the given location within that period.

3.7 System requirements

The system requires adequate resources to enable it execute seamlessly. Both the hardware and software should match the system demands to reduce execution time.

They include:

I. An operating system which is the platform that supports all other softwares. In this case the OS used is windows 8.1
II. Google maps API.
III. Google geo code API.
IV. Xampp server.
V. Boot strap framework.
VI. A computer with a Ram of 3GB and above.

CHAPTER 4: SYSTEM ARCHITECTURE AND DESIGN

The system is has two major parts namely: The client side and the server side. A web browser represents the client side which enables users interact with the system. The server side is made up of the tweets database and a server.

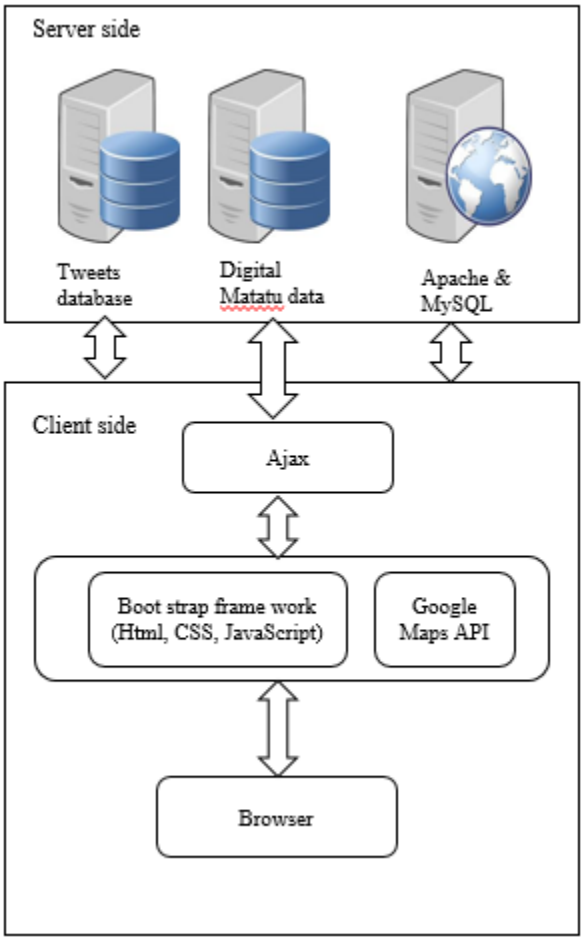


Figure 8: System architecture

(Source: Research)

The tweets are saved in the tweets database which contains all the information about the original tweets, time it was sent and the status associated with that tweet. Most of the users have not activated the geo locator on their devices hence the location mentioned in the tweet is picked to be the user's current location.

4.1 Client side

The client side is made up of:

4.1.1 User interface.

Users interact with the system via the user interface. A web browser acts as the bridge between the user and the system. By use of HTML, JavaScript and CSS the user interface is developed. The bootstrap framework which combines HTML, CSS and JavaScript has been used. To enable display of query results, a Google API is added to enable the system pull out Google maps on which the points of interest is mapped on.

4.1.2 Google map API

Google maps API provide a way for developers to embed maps to web pages thus enabling users to easily visualize information. The API loads the Map on which the location markers are displayed and users can manipulate the marker by zooming in and out to get a clearer picture. Its provided freely by the Google company

4.1.3 Google geo-code API

Google geo code API helps generate the Geo coordinates i.e. the latitude and longitude of a given location. The *GetLatitude* and *GetLongitude* functions are used to return the Latitude and longitude of an address respectively. The output is used by Google maps API to display a marker of the point of interest to the user.

4.2 Server side

The server side has the following components

4.2.1 Tweets database

The database contains all the tweets received from *ma3route* application. A SQL database is used in this case due to its availability and it's also dynamic. A tweet is stored in its original format in the database containing all the details which include time, date, tweet ID, traffic status associated with it, Senders username and Geo-coordinates if the user has activated geo location on their device.

4.2.2 Xammp server

Xammp server comes as a single package that comprises of most web technology development tools Dvorski (2007). Due to its portability and light size, it's mostly preferred when developing

Php and MySql applications. In this prototype, Xampp is used as the local server. It's freely available for download at http://www.apachefriends.org/ .

4.3 System Design

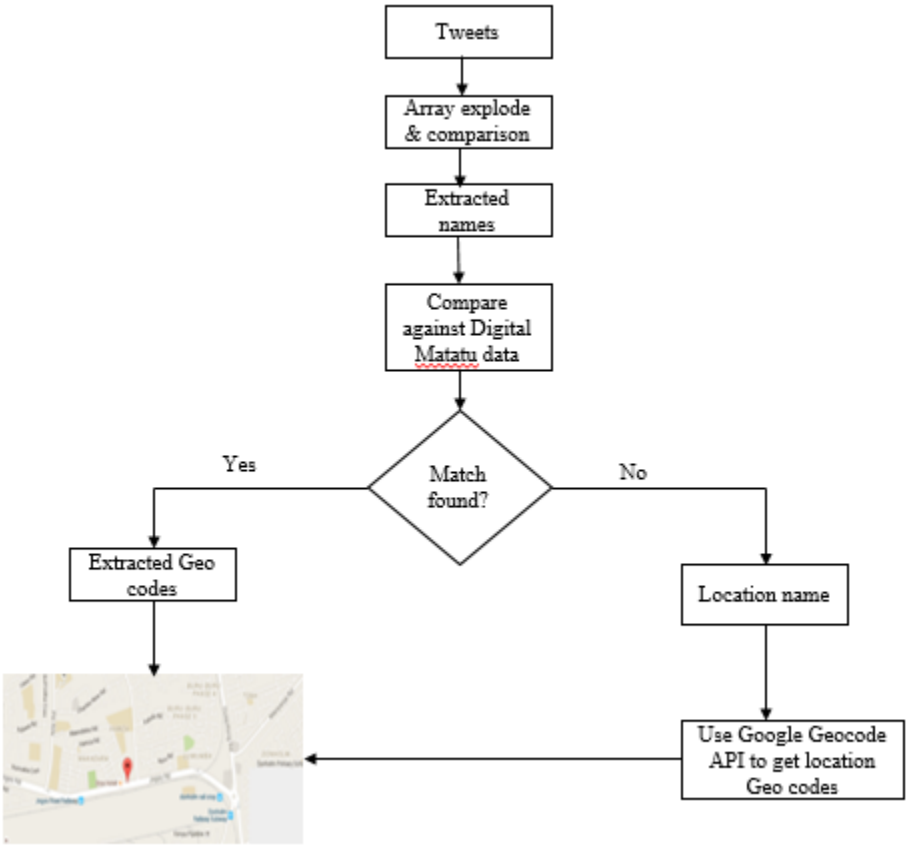


Figure 9: System Design diagram

(Source: Research)

The implementation of the system is done in several phases. The initial step is to save the tweets in the database. The original tweets are received in comma separated value (CSV) format and converted to SQL statements by the SQL 'import from CSV' function available in MySQL database. The tweet is exploded and string comparison is performed depending on the query. The result is matched against local database of Digital Matatu data and if a match is found, the result

which is a Latitude and Longitude are plotted on Google maps else the result is sent to Geo code API to get its Geo coordinates then plotted on Google maps.

A68 249875

	A	B	C	D	E
1	ification_id	description	isactive	severity	date
2	249993	the miracle that was Langata road this morning(8 ish)..kudos to the ones controlling traffic today http://t.co/tyY9GRhnzY	Y	General Info	2015-05-20 23:57:02.178
3	249990	what's wrong today?? No traffic ...did like 15mins from gigiri to Riverside at 9..	Y	General Info	2015-05-20 23:53:05.198
4	249989	Giving way means sometimes means veering off the road and that endangers life of cyclists and pedestrians.	Y	General Info	2015-05-20 23:52:59.828
5	249987	Thika rd,Muranga rd same 2 Forest rd is very clear mpaka tao	Y	Clear	2015-05-20 23:51:10.605
6	249986	muhoho rd http://t.co/Q8UH3isaGf	Y	Bumper To Bumper	2015-05-20 23:49:24.055
7	249985	forest road from Pangani junction baad	Y	Bumper To Bumper	2015-05-20 23:46:31.226
8	249984	westlands is clear	Y	Clear	2015-05-20 23:44:52.755
9	249983	akila http://t.co/LtPttvmTgP	Y	General Info	2015-05-20 23:44:49.915
10	249982	No traffic at Globe	Y	Clear	2015-05-20 23:42:46.17
11	249979	Thika Road absolutely clear	Y	Clear	2015-05-20 23:40:09.157
12	249976	The police officers at Nyayo Stadium should be held fully to account, should handcarts cause accidents whilst on wrong way on Mombasa road	Y	General Info	2015-05-20 23:36:05.566
13	249975	dar slm rd not bad http://t.co/0l6cZqjstN	Y	General Info	2015-05-20 23:35:05.83
14	249974	jogoo Road to town is smooth	Y	Clear	2015-05-20 23:33:38.872
15	249973	today from JKIA to buru at amd 07:50hrs took me like 30min to buru	Y	General Info	2015-05-20 23:33:30.364
16	249972	enterprise rd both sides http://t.co/SYeYElXzKM	Y	General Info	2015-05-20 23:33:27.745
17	249971	Traffic police officers at Nyayo Stadium roundabout brazenly allow handcart pullers onto the WRONG WAY of Mombasa Road! @JBoinnet !!!!	Y	General Info	2015-05-20 23:32:56.918
18	249970	jogoo rd no nyweeeee from likoni rd http://t.co/Z1VtlWDDe3	Y	Clear	2015-05-20 23:31:54.064

Figure 10: Sample Ma3Route data

(Source: Ma3Route)

Clipboard | Font | Alignment | Number | Styles | Cells | Editing

I21

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	stop_id	stop_code	stop_name	stop_desc	stop_lat	stop_lon	zone_id	stop_url	location_t	parent_station											
2	0005AMB		Ambassadeur	D	-1.285963	36.826048			1												
3	0006BSS		Bus Station	D	-1.287191	36.828881			1												
4	0007COM		Commercial	D	-1.284282	36.826188			1												
5	0512BST		Bus Station	D	-1.287281	36.828863				0006BSS											
6	0512BSN		Bus Station	D	-1.287635	36.829215				0006BSS											
7	0512BSL		Bus station	D	-1.287229	36.829477				0006BSS											
8	0512CML		Commercial Imara	D	-1.284306	36.826257				0007COM											
9	0512CMC		Commercial	D	-1.283947	36.826437				0007COM											
10	0512AMU		Ambassadeur	D	-1.285315	36.825615															
11	0512ESC		Easy Coach/Uchumi	ND	-1.289218	36.828251															
12	0512EAS		Easy Coach/Uchumi	D	-1.288854	36.828734															
13	0001RLW		Railways Terminus	D	-1.290884	36.828242			1												
14	0512RLW		Railways	D	-1.290352	36.828268				0001RLW											
15	0513KWR		Kware	D	-1.396812	36.75024															
16	0412KNC		Kencom	D	-1.28589	36.82429															
17	6050GPO		GPO	D	-1.28604	36.818211															

Figure 11: Sample Digital Matatu data

Source: GTFS exchange website:[4]

[4] GTFS Exchange Website: http://www.gtfs-data-exchange.com/agency/university-of-nairobi-c4dlab/

The tweets in the database have several fields but the key one includes description which is the actual tweet, severity which corresponds to traffic status, date and time.

4.4 Algorithm implementation

When a user issues a query, the location is the keyword that is used to filter the tweets. The time frame which the user has specified defines the extent of search. All the tweets which contain the keyword are considered. Each tweet is handled at a time by segmenting each word and comparing it to the search word. If a match is found in a tweet within the given time frame, the corresponding status is picked. The sum of similar status is calculated and the mode status is picked as the traffic status of that particular location for that given period and is displayed on Google maps.

If the given location does not match any word in the tweets within the specified time, a location marker is placed on Google maps with a no update yet message.

CHAPTER 5: IMPLEMENTATION

5.1 System overview

Once the system is designed, it is tested using data from *Ma3route* application and Digital Matatu. On loading the application a map for Nairobi area is displayed which is fetched from Google maps using Google map API.

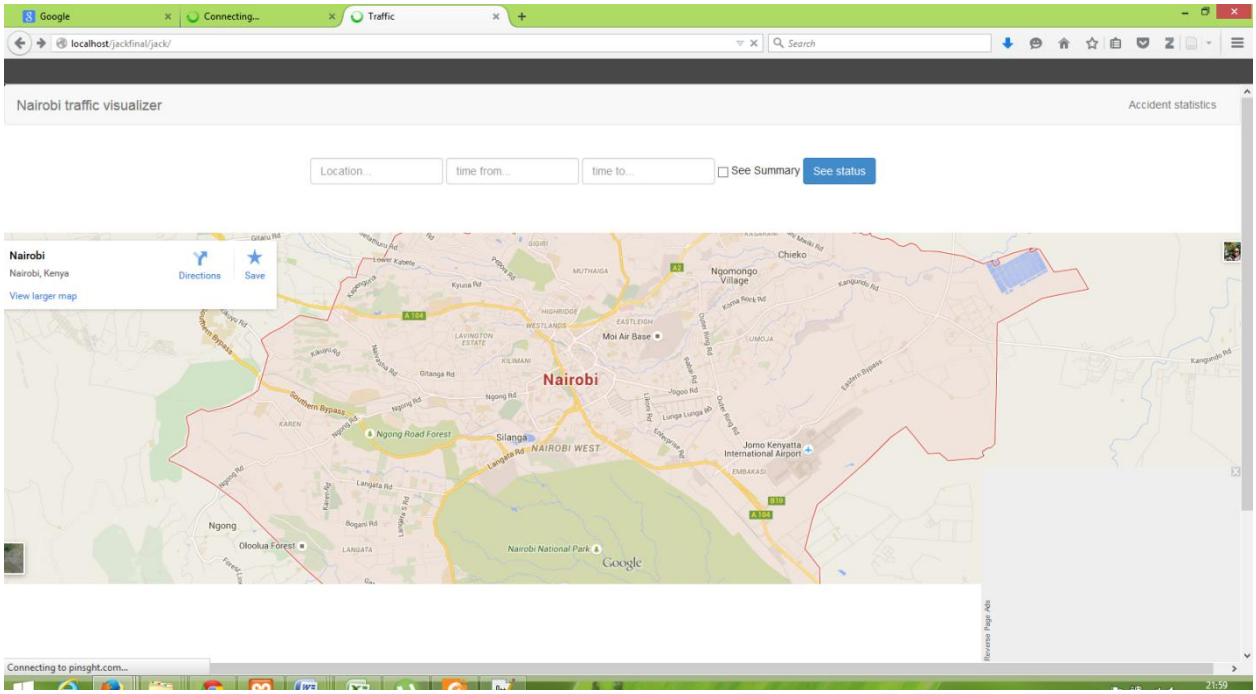


Figure 12: Nairobi Area Map

The home page contains a text box where the user types in the search location, time from, time to, a check box for status then a command button for see status. Another main menu on this page is for accident statistics.

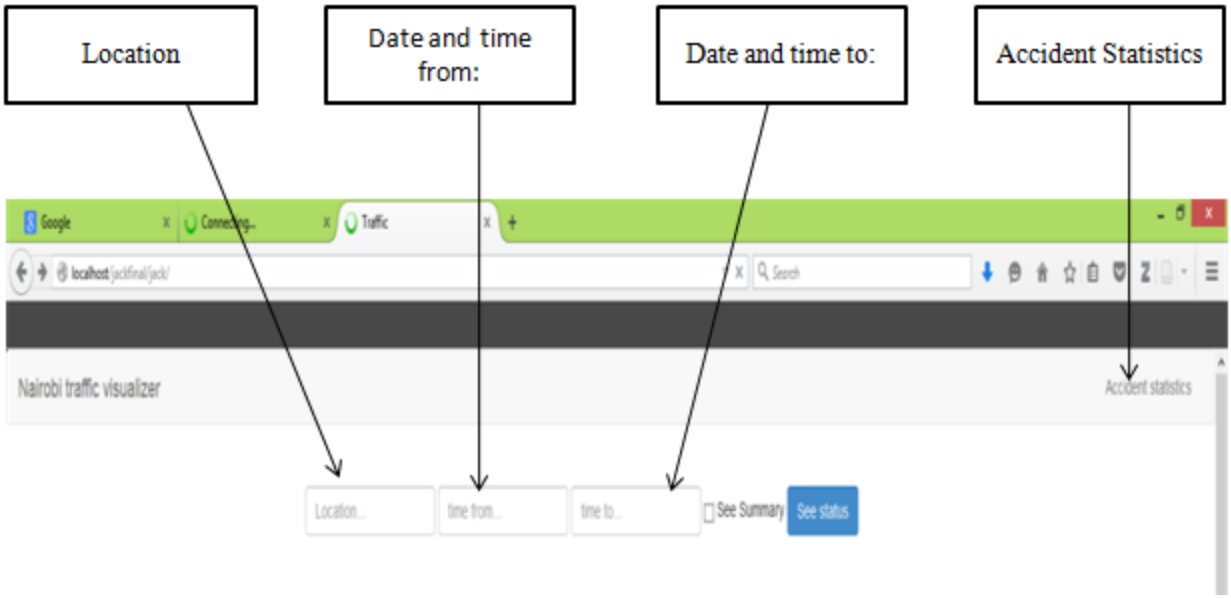


Figure 13: Nairobi traffic visualizer

The details on the location, time from and time to text boxes can be used to invoke two functionalities.

a) ‘See status`

The tweets in the database are filtered according to the key word i.e. location then all tweets in which the keyword appears are selected. Next, the period defined by the user is used to filter the tweets and only those within that time frame and with the given location are chosen. Each tweet is associated with a status mode i.e. Bumper to Bumper, Moderate or Clear. The mode status in the selected tweets is concluded to be the traffic status of that particular location within the indicated time frame. If no mentions about that location are received within the specified period, the system returns a feedback message to the user indicating that no updates have been found.

b) See summary see status

The tweets within the indicated time frame for that location are selected. The status of all the tweets associated with a particular status are aggregated and computed as a ratio of the total tweets and represented on a pie chart and a line graph

5.2 System tests and results

5.2.1 Traffic status and summary

Case 1

User types location name, selects date and time from the calendar then clicks the command button ‘See status’. If there are tweets for the given location within the specified time period, the user gets a feedback which indicates the mode status of the particular location. The date picker function enables the user to pick the date and time in a standard method rather than typing to avoid date format errors.

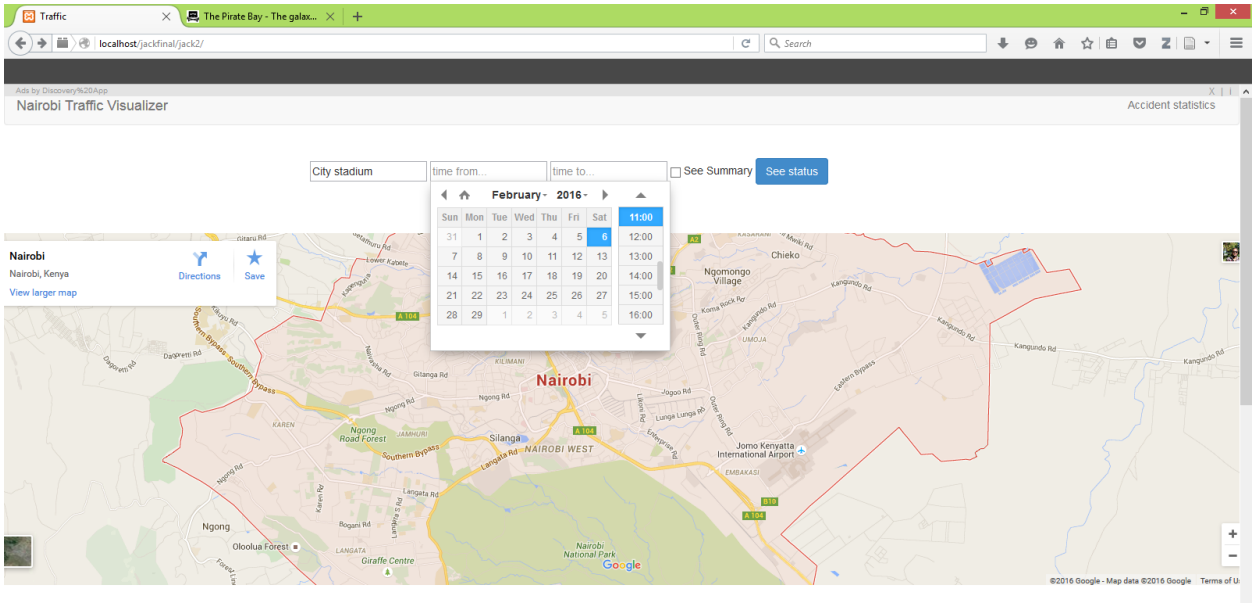


Figure 14: Use case 1 – Date and time from

Next is to select the end of the desired time frame i.e. `Date to’

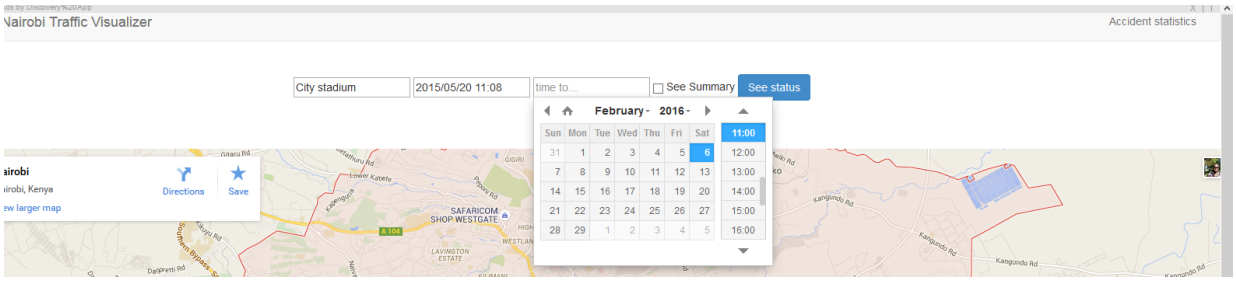


Figure 15: Use Case 1 - Date and Time to

The result is a marker that is pinned to the location on the Google maps and upon clicking it, it shows the mode traffic status within the specified period.

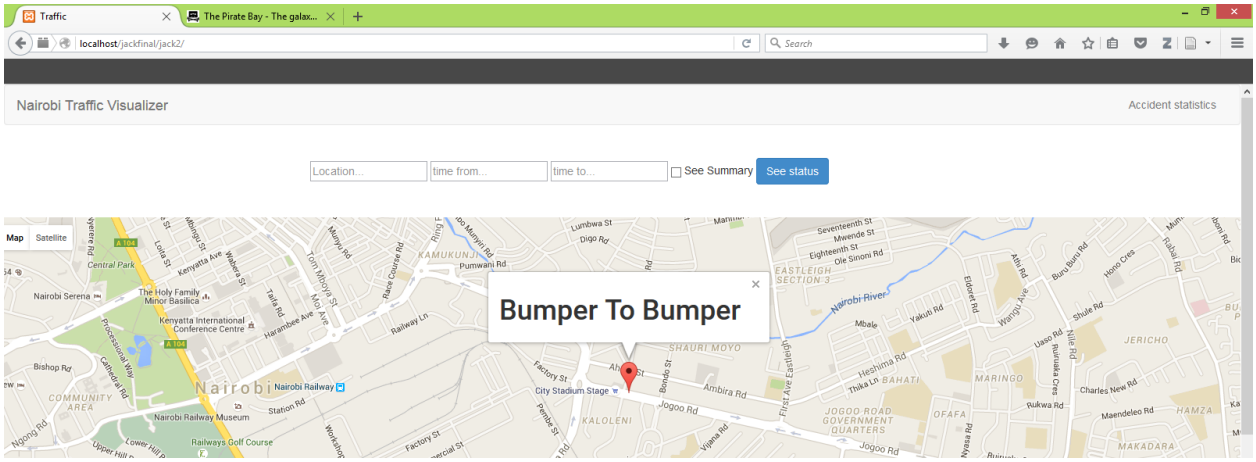


Figure 16: Use Case 1 Results

Case 2

User types location name, selects the ‘date from’ and `date to’ then command button ‘See Status’. If there are no tweets for the given location within the specified time period, the user gets a feedback that no updates are available for that particular location.

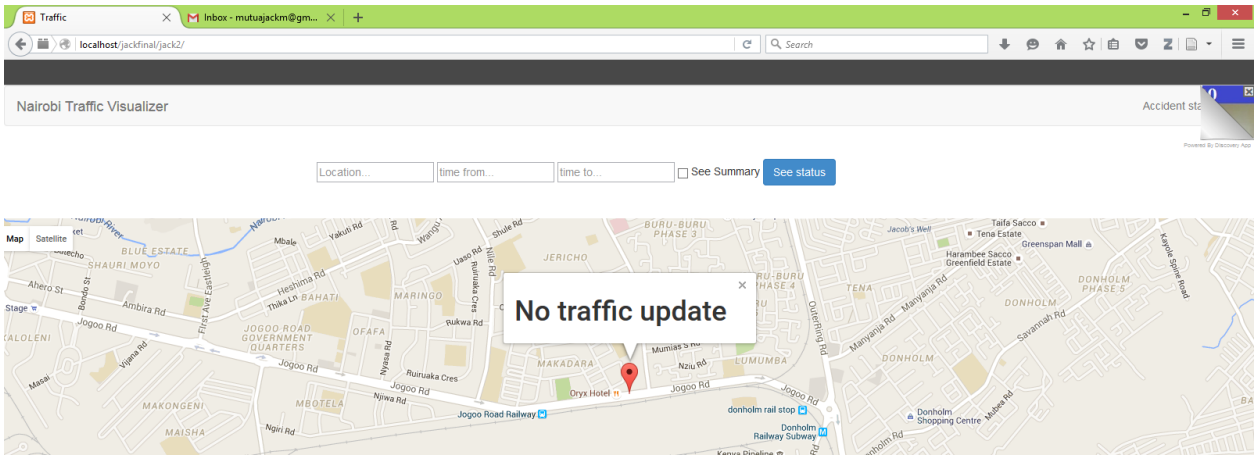


Figure 17: Use case 2 Results

Case 3

User types location name, time from and time to checks the ‘See summary ‘box then clicks the command button ‘See Status’. The result is pie chart showing the distribution of tweets according to status the tweets of the location within the specified period.

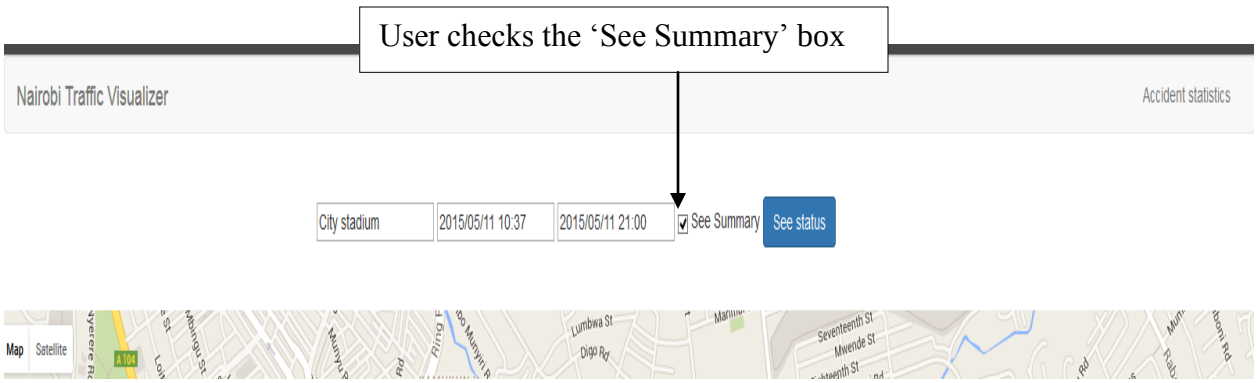


Figure 18: Use case 3

The result is a pie chart showing the distribution of traffic status as per the received tweets within the specified period for the point of interest. In this case all status are Bumper to Bumper.

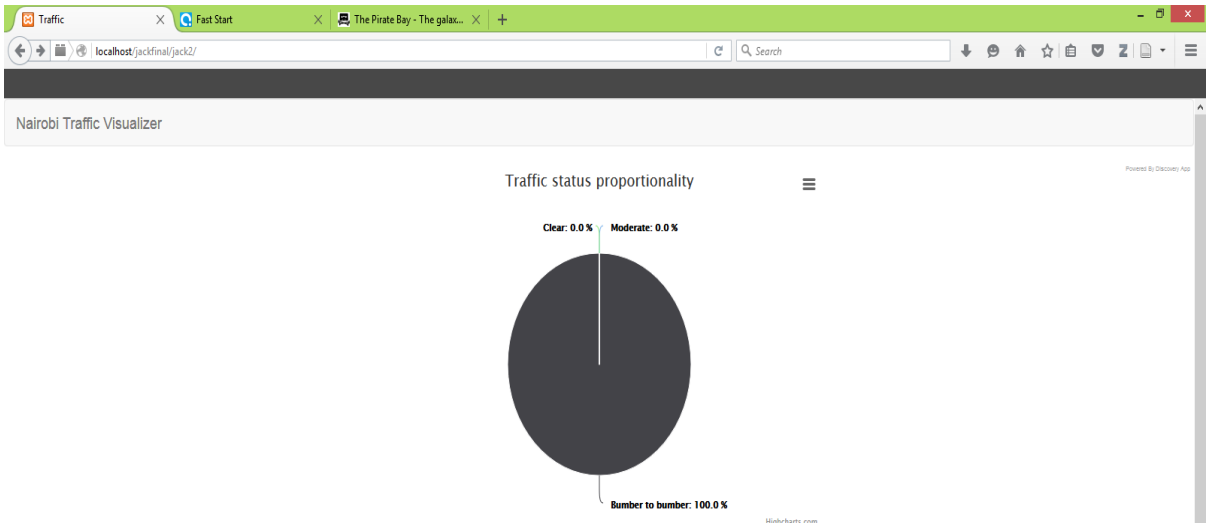


Figure 19: Use case 3 Results

The same result is represented in line graph showing the status on different levels. The key provided helps the user to interpret the results.

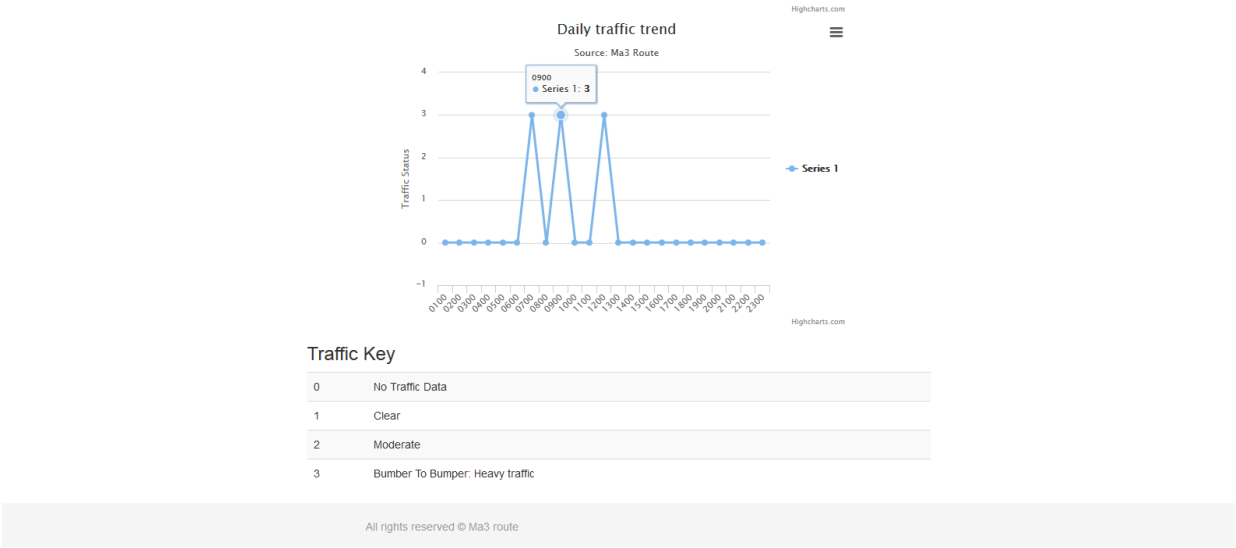


Figure 20: Graphical results output

5.2.2 Accident statistics

A user selects a time period for the system to populate the accidents report, the types of vehicles involved and locations whereby the number of accidents on that particular place is indicated. The system uses key words like “Accident, Crush, Collision, Ajali” to traverse the database and select tweets in which these key words appear. If the search word matches a tweet, the tweet is

selected and exploded, words compared to a local dictionary for location and vehicle type as defined. The location mentioned on the tweet is picked and also the type of vehicle if the person who posted the message has indentified it.

The number of mentions of a specific location within the given time period with either of the key words appearing in each tweet is computed to be the number of accidents at that point within that time.

User keys in time period

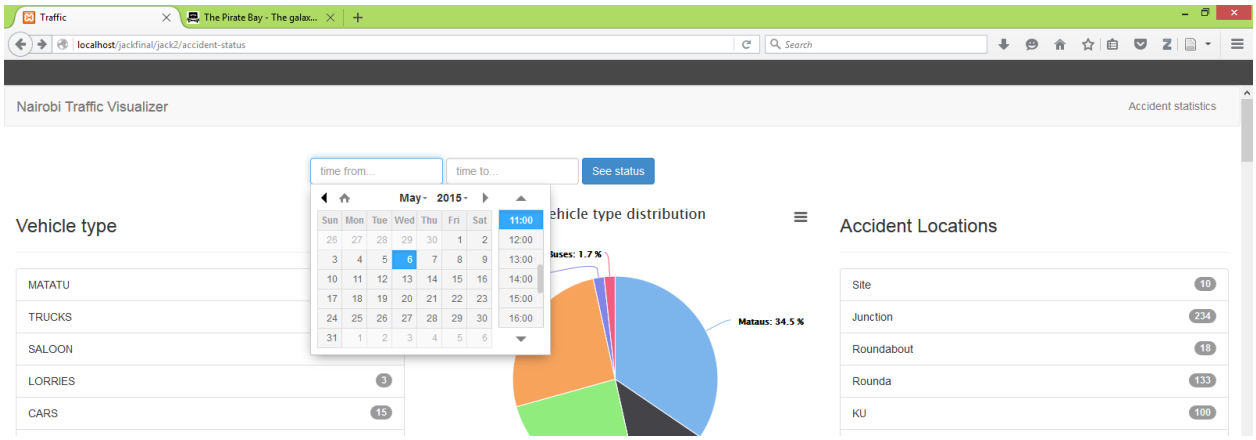


Figure 21: Accident statistics

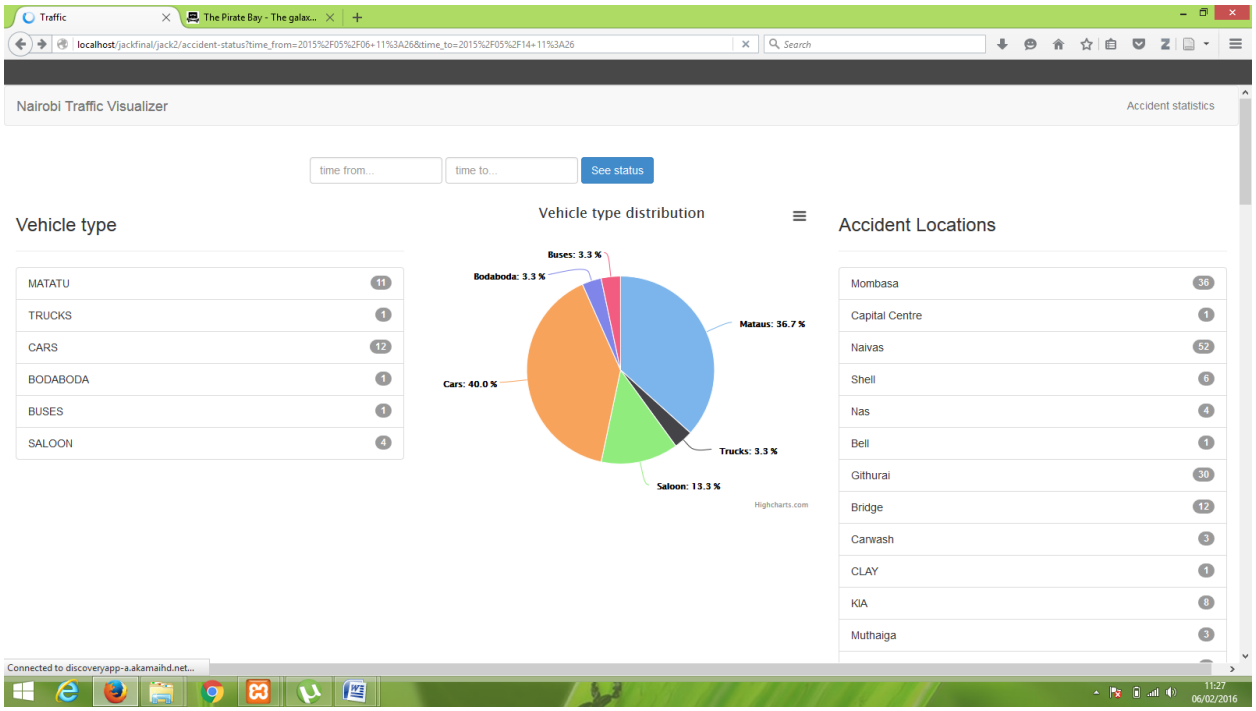


Figure 22: Accident statistics Query results

CHAPTER SIX: FINDINGS AND CONCLUSION

6.1 Findings

6.1.1Road status

When data from a particular location is analyzed and visualized for a given time it is easy to establish the peak periods it and compare traffic status distribution. For the below location i.e. Donholm, from the data received congestion starts from 3:00am to 5:00 am then from 6:00 am to 9:00 am and then again from 19:00hrs to 22:00hrs. From 22:00hrs to 23:00 the received data indicates that the traffic conditions were moderate. The periods denoted by Bumper to Bumper are picked as peak time and moderate. The rest of the day could have different traffic conditions but no data was received.

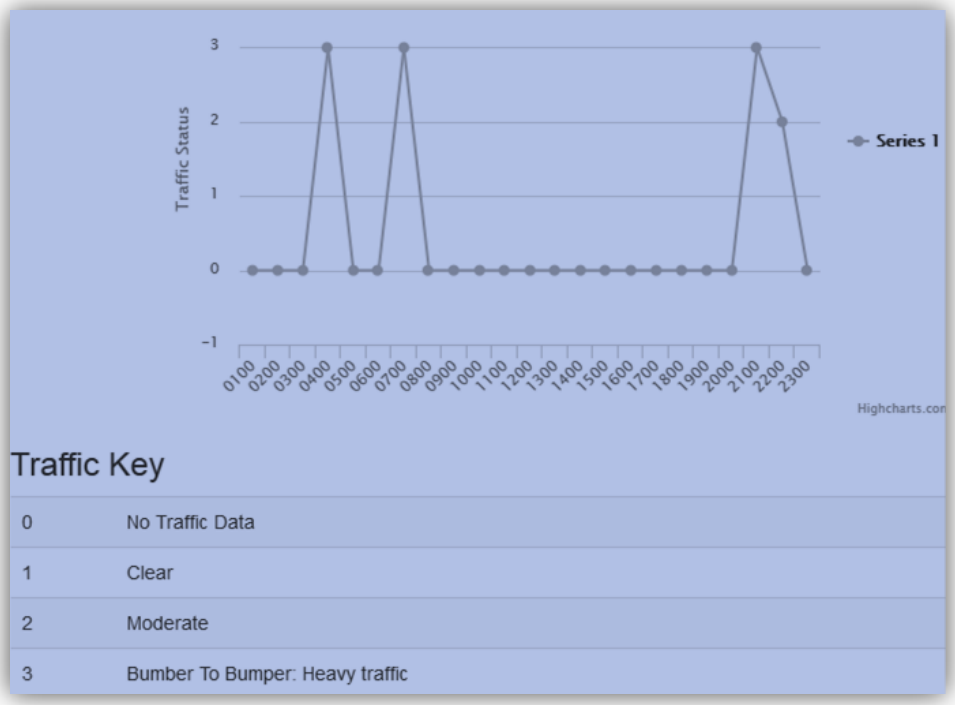


Figure 23: Traffic status distribution for Donholm

Source: Research

The same distribution is shown in figure 24 but in terms of percentage

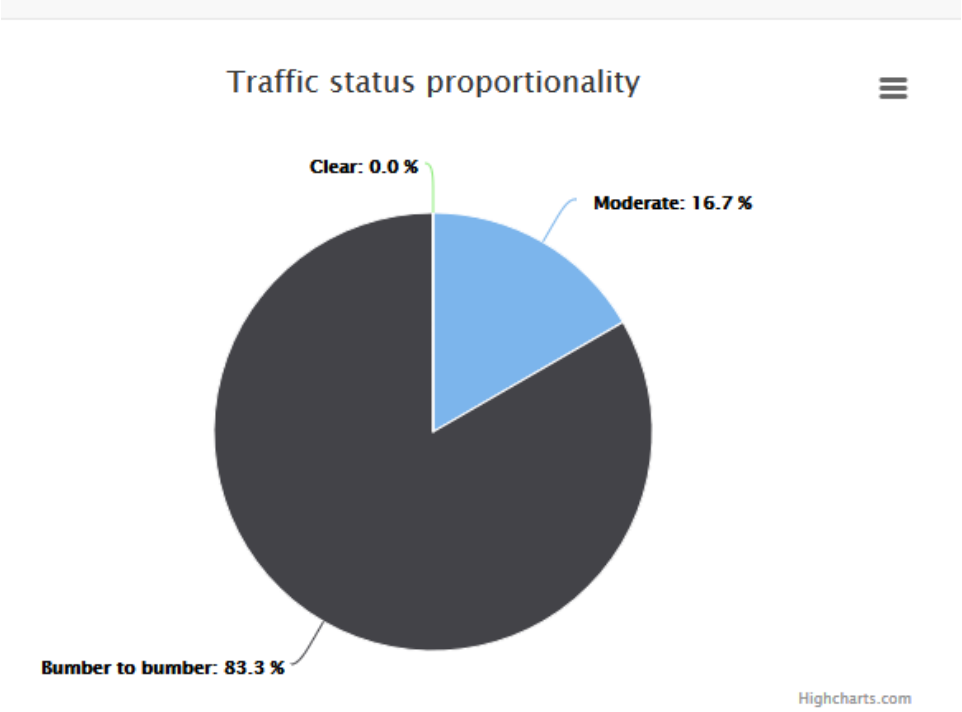


Figure 24: Status proportionality for Donholm

Source: Research

6.2 Accidents statistics

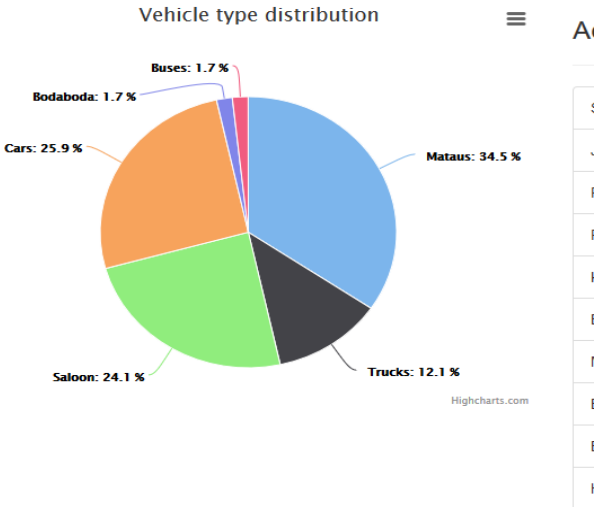


Figure 25: Percentage of Vehicle types in accidents

Tweets for twenty days show that Matatus were the most mentioned in accidents tweets. Using key words like ‘Ajali’, ‘Crush’ and ‘Accident’ to filter through the tweets and also get locations. From the data received we can conclude that Matatus are involved in the highest number of accidents.

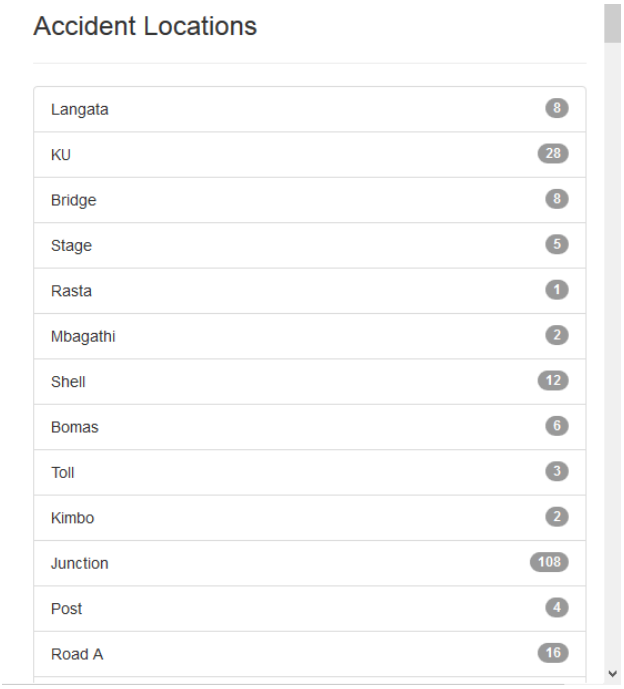


Figure 26: Accident Locations

The locations in figure 26 were mentioned in the accident tweets within the period the data was collected. The number of times mentioned is taken as the number of accidents for that particular location.

6.2 Conclusion

Social media has provided a platform for users to share information in regards to different subjects. Twitter has been widely used in providing real time information such as earthquakes, prediction of outcomes depending on trending topics and traffic updates. A lot of data is used posted on @ma3route twitter handle which when mined and visualized can show general trend of how people feel about traffic. Visualization makes it easier for people to understand and interpret facts found in data. Social media can be used as a source of reliable cheap real time.

Ma3route data has been categorized depending on the status associated with the tweet. There are posts about accidents, enquiries on route status, general information and reporting on route status which is most common. The output represented using both the pie chart and line graph indicate that most of the tweets posted in the morning and evening have most status as Bumper to Bumper which is a clear indication that traffic congestion during these periods is widely experienced. No standard format is used while posting the tweets therefore processing the data is a challenge. Actual checks on the roads stops such as Hamza, City stadium and Ladhies road against the data posted on *Ma3route* reveals that tweets regarding these areas match the message on the tweet. An average lag of five minutes was observed from when the actual check time to the when the data was available on ma3route twitter handle.

Google directions display the traffic conditions on a route from start point to end point of a journey. When compared against *Ma3route* data, areas denoted as congested represented by the red colour matched tweets regarding that particular area.

Digital Matatu data which serves as the local dictionary for locations having been validated against Google maps before upload improves accuracy on locating areas mentioned on the tweet. Its open source nature aids in reducing cost input in the project.

RAID evolutionary prototyping methodology ensured refining of previous work without discarding previous work. Use of open source software such as bootstrap framework made prototyping easier and quicker to customize which translates to its adoption in different environments.

6.2 Limitations and challenges of the study

1) Limited access to data forced validation of prototype to be done with available data

2) Data available on *Ma3route* does not relate to traffic only therefore a lot of processing is required to filter the required information.

3) Tweets are not structured in a particular way therefore most of the existing tools could not classify the tweets relating to traffic and use of slang language complicates identification of nouns which represent locations in this study

4) Direction is not specified therefore it is difficult to determine whether traffic is inwards or outwards.

5) When user refers to a street or road, assuming the entire street is experiencing that traffic status could be incorrect given that it could be only on a particular segment on that road.

6) Data correctness is subject to the user accuracy. A user can post while they are far away from the point of interest thus increasing lag in time.

7) Most users have not activated Geo location feature on their devices therefore location is assumed to be the place mentioned in the tweet.

6.3 Future work

The study relies on *Ma3route*, Digital Matatu and nrb.city data only. Comparing data from emerging sources such as @kenyatraffic, @myroadtraffic and @RoadAlertsKE will make the data more comprehensive. The system should be enhanced to fetch data directly from *Ma3route* servers to make data visualization real time.

REFERENCES

1. Asur, S., &Huberman, B. A. (2010). Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on* (Vol. 1, pp. 492-499). IEEE.
2. Beaudouin-Lafon, M., & Mackay, W. E. (2003). Prototyping tools and techniques. *Human Computer Interaction—Development Process*, 122-142.
3. Boulos, M. N. K., Resch, B., Crowley, D. N., Breslin, J. G., Sohn, G., Burtner, R., ...& Chuang, K. Y. S. (2011). Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, OGC standards and application examples. *International journal of health geographics*, *10*(1), 67.
4. Budi, I., &Bressan, S. (2003). Association rules mining for name entity recognition.
5. Carvaloh, S.F.L (2010) Real time sensing of traffic information in twitter messages.
6. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., &Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, *12*, 2493-2537.
7. Doan, A., Franklin, M. J., Kossmann, D., &Kraska, T. (2011). Crowdsourcing applications and platforms: A data management perspective. *Proceedings of the VLDB Endowment*, *4*(12), 1508-1509.
8. Dvorski, D. D. (2007). Installing, configuring, and developing with Xampp. *Skills Canada*.
9. Endarnoto, S.K., Pradipeta, S., Nugroho, A.S. and Purnama, J., 2011, July. Traffic condition information extraction & visualization from social media twitter for android mobile application. In Electrical Engineering and Informatics (ICEEI), 2011 International Conference on (pp. 1-4). IEEE.
10. Elsafoury, F,A,. (2013) Monitoring urban traffic status using twitter messages.
11. Jakimavičius, M. (1999). Traffic flows analysis and visualization based on data from an advanced Vilnius traveller's information system. *Indicator*, *2005*, 2012.

12. Gordon, V. S., & Bieman, J. M. (1995). Rapid prototyping: lessons learned. *IEEE software*, *12*(1), 85-95.
13. Guduru, N. (2006). *Text mining with support vector machines and non-negative matrix factorization algorithms* (Doctoral dissertation, University of Rhode Island).
14. Klopp, J., Williams, S., Waiganjo, P., Orwa, D. and White, A., 2015. Leveraging Cellphones for Wayfinding and Journey Planning in Semi-formal Bus Systems: Lessons from Digital Matatus in Nairobi. In Planning Support Systems and Smart Cities (pp. 227-241). Springer International Publishing.
15. Kosala, R., &Adi, E. (2012). Harvesting real time traffic information from twitter. *Procedia Engineering*, *50*, 1-11.
16. Kumar P., Singh V., and Reddy D., (2005) Advanced Traveler Information System for Hyderabad City, *IEEE Trans. Intell. Transp.Syst.*, vol. 6, no. 1.
17. Lécué, F., Tucker, R., Bicer, V., Tommasi, P., Tallevi-Diotallevi, S., &Sbodio, M. (2014). Predicting severity of road traffic congestion using semantic web technologies. In *The Semantic Web: Trends and Challenges* (pp. 611-627). Springer International Publishing.
18. Liu, K., Deng, K., Ding, Z., Li, M., & Zhou, X. (2009). Moir/mt: Monitoring large-scale road network traffic in real-time. *Proceedings of the VLDB Endowment*, *2*(2), 1538-1541.
19. Li, C., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A., & Lee, B. S. (2012). Twiner: named entity recognition in targeted twitter stream. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval* (pp. 721-730). ACM.
20. Huber, W., Lädke, M., & Ogger, R. (1999, November). Extended floating-car data for the acquisition of traffic information. In *Proceedings of the 6th World congress on intelligent transport systems* (pp. 1-9).
21. Misra, A., Gooze, A., Watkins, K., Asad, M., & Le Dantec, C. A. (2014). Crowdsourcing and Its Application to Transportation Data Collection and Management. *Transportation Research Record: Journal of the Transportation Research Board*, *2414*(1), 1-8.
22. Muthiah, G., Prashant, S., Pushpa, V., Natarajan, J., Jhunjhunwala, A., &Waidyanatha, N. (2011). The use of mobile phone as a tool for capturing patient data in southern rural Tamil Nadu, India. *Journal of Health Informatics in Developing Countries*, *5*(2).

23. Pack, M. L., Weisberg, P., & Bista, S. (2005). Four-dimensional interactive visualization system for transportation management and traveler information. *Transportation Research Record: Journal of the Transportation Research Board*, *1937*(1), 152-158.

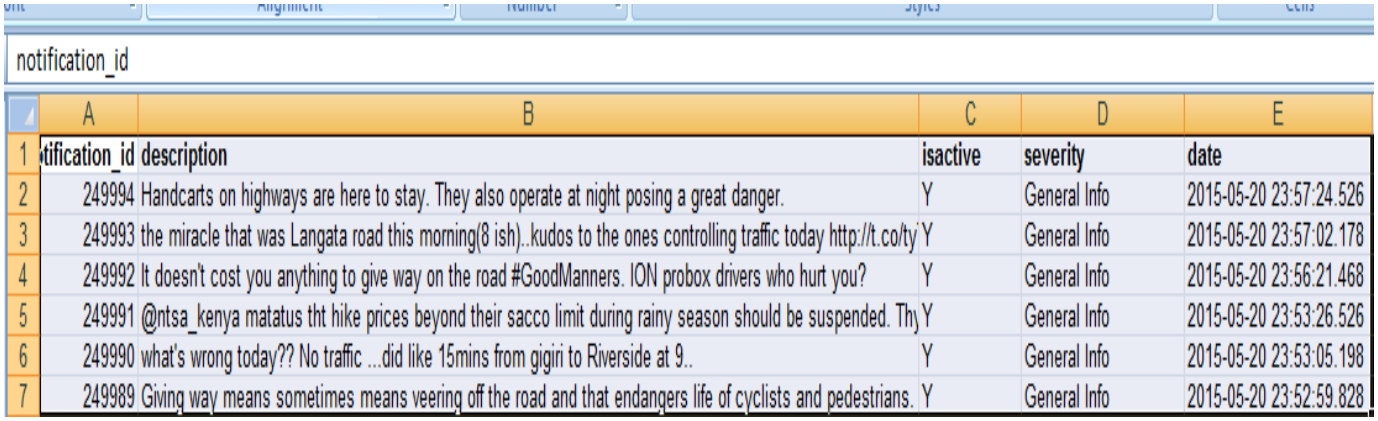
24. Portillo, D. (2008). *Automated Vehicle Location using Global Positioning Systems for First Responders*. Air force academy Colorado springs co inst for information technology applications.
25. Preotiuc-Pietro, D., Samangooei, S., Cohn, T., Gibbins, N., &Niranjan, M. (2012, July). Trendminer: An architecture for real time analysis of social media text. In *Proceedings of the workshop on real-time analysis and mining of social streams*.
26. Ribeiro, A. I. J. T., Silva, T. H., Duarte-Figueiredo, F., &Loureiro, A. A. (2014). Studying traffic conditions by analyzing foursquare and instagram data. In *Proceedings of the 11th ACM symposium on Performance evaluation of wireless ad hoc, sensor, & ubiquitous networks* (pp. 17-24). ACM.
27. Ritter, A., Clark, S., &Etzioni, O. (2011). Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1524-1534). Association for Computational Linguistics.
28. Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web* (pp. 851-860). ACM.
29. Santani, D., Njuguna, J., Bills, T., Bryant, A.W., Bryant, R., Ledgard, J. and Gatica-Perez, D., 2015, August. CommuniSense: Crowdsourcing Road Hazards in Nairobi. In Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services (pp. 445-456). ACM.
30. Statista (2016) Number of social network users worldwide from 2010 to 2019 (in billions) [online] Availabel at: www.statista.com/statistics/278414/**number**-of-worldwide-**social-network-users/** (accessed 04 July 2016)
31. String explode available from: http://php.net/manual/en/function.explode.php. (Accesed 08 August 2015).

32. Tao, S., Manolopoulos, V., Rodriguez, S., &Rusu, A. (2012). Real-time urban traffic state estimation with A-GPS mobile phones as probes. *Journal of Transportation Technologies*, *2*(01), 22.
33. Walton, S., Chen, M., & Ebert, D. LiveLayer: Real-time Traffic Video Visualisation on Geographical Maps.
34. Yaghini, M., Bourouni, A., & Amiri, R. H. (2009). A framework for selection of information systems development methodologies. *Computer and Information Science*, *2*(1), p3.

APPENDIX

Sample Tweets

General info tweets

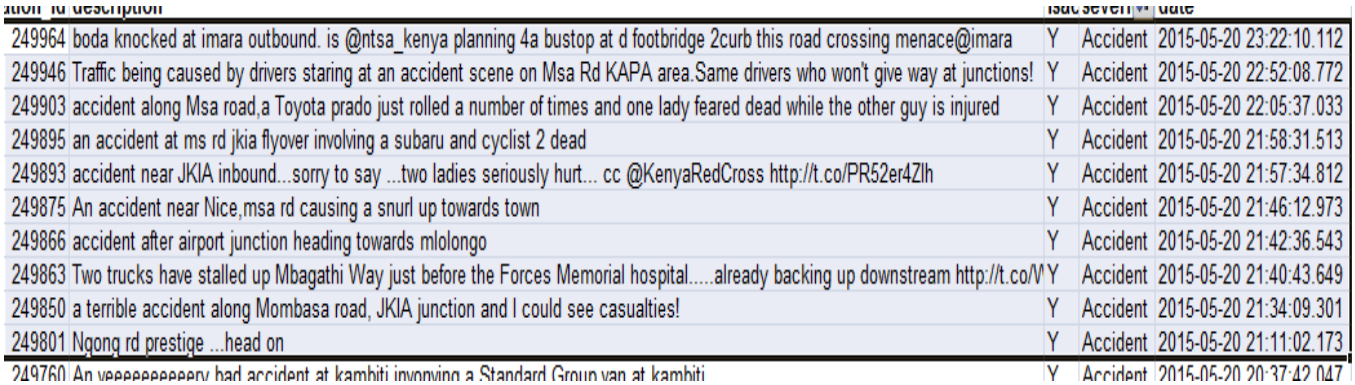


notification_id

	A	B	C	D	E
1	tification_id	description	isactive	severity	date
2	249994	Handcarts on highways are here to stay. They also operate at night posing a great danger.	Y	General Info	2015-05-20 23:57:24.526
3	249993	the miracle that was Langata road this morning(8 ish)..kudos to the ones controlling traffic today http://t.co/ty	Y	General Info	2015-05-20 23:57:02.178
4	249992	It doesn't cost you anything to give way on the road #GoodManners. ION probox drivers who hurt you?	Y	General Info	2015-05-20 23:56:21.468
5	249991	@ntsa_kenya matatus tht hike prices beyond their sacco limit during rainy season should be suspended. Th	Y	General Info	2015-05-20 23:53:26.526
6	249990	what's wrong today?? No traffic ...did like 15mins from gigiri to Riverside at 9..	Y	General Info	2015-05-20 23:53:05.198
7	249989	Giving way means sometimes means veering off the road and that endangers life of cyclists and pedestrians.	Y	General Info	2015-05-20 23:52:59.828

Road users' report a lot of happenings not necessarily traffic updates on ma3route

Accident tweets



249964	boda knocked at imara outbound. is @ntsa_kenya planning 4a bustop at d footbridge 2curb this road crossing menace@imara	Y	Accident	2015-05-20 23:22:10.112
249946	Traffic being caused by drivers staring at an accident scene on Msa Rd KAPA area.Same drivers who won't give way at junctions!	Y	Accident	2015-05-20 22:52:08.772
249903	accident along Msa road,a Toyota prado just rolled a number of times and one lady feared dead while the other guy is injured	Y	Accident	2015-05-20 22:05:37.033
249895	an accident at ms rd jkia flyover involving a subaru and cyclist 2 dead	Y	Accident	2015-05-20 21:58:31.513
249893	accident near JKIA inbound...sorry to say ...two ladies seriously hurt... cc @KenyaRedCross http://t.co/PR52er4Zlh	Y	Accident	2015-05-20 21:57:34.812
249875	An accident near Nice,msa rd causing a snurl up towards town	Y	Accident	2015-05-20 21:46:12.973
249866	accident after airport junction heading towards mlolongo	Y	Accident	2015-05-20 21:42:36.543
249863	Two trucks have stalled up Mbagathi Way just before the Forces Memorial hospital.....already backing up downstream http://t.co/V	Y	Accident	2015-05-20 21:40:43.649
249850	a terrible accident along Mombasa road, JKIA junction and I could see casualties!	Y	Accident	2015-05-20 21:34:09.301
249801	Ngong rd prestige ...head on	Y	Accident	2015-05-20 21:11:02.173
249760	An yeeeeeeeeery bad accident at kambiti involving a Standard Group van at kambiti	Y	Accident	2015-05-20 20:37:42.047

Sample codes

Loading Nairobi map

```
function initialize() {
        var myLatlng = new google.maps.LatLng({{$lat}}, {{$long}});
            var mapOptions = {
            zoom: 15,
                center: myLatlng
```

```
                };
                var map = new google.maps.Map(document.getElementById('map-canvas'), mapOptions);
                var contentString = "<div id='content'><h1><?= $status ?></h1></div>";
                var infowindow = new google.maps.InfoWindow({
                content: contentString
                });
                var marker = new google.maps.Marker({
                position: myLatlng,
                        map: map,
                        title: '<?= $name ?>'
                });
                google.maps.event.addListener(marker, 'click', function() {
                infowindow.open(map, marker);
                });
            }
        google.maps.event.addDomListener(window, 'load', initialize);</script>
    <style>
        html { position: relative; min-height: 100%; }
        body {margin-bottom: 60px; }
        #footer {position: absolute; bottom: 0; width: 100%; height: 60px; background-color: #f5f5f5; }
        /* Custom page CSS
        ------------------------------- */
        /* Not required for template or sticky footer method. */
        .container { width: auto; max-width: 680px; padding: 0 15px; }
        .container .text-muted { margin: 20px 0; }
    </style>
  </head>
Home Page
```

```
<div class="row">
        <div style="width: 800px; margin: 0 auto; padding-top: 20px; padding-bottom: 20px;">
            <form class="form-inline" method='post'>
                <div class="form-group">
                    <label class="sr-only">Location Name</label>
                    <input type="text" name='location' id="exampleInputEmail2" placeholder="Location...">
                </div>
                <div class="form-group">
                    <label class="sr-only">Time From</label>
                    <input type="text" name='time_from'  id="date_x" placeholder="time from...">
                </div>
                <div class="form-group">
                    <label class="sr-only">Time To</label>
                    <input type="text" id="date_y" name='time_to'  placeholder="time to...">
                </div>
                <div class="form-group">
                    <label class="sr-only" for="exampleInputPassword2">See summary</label>
                    <input type="checkbox" class="form-control" name="summary" /> See Summary
                </div>
                <input type="hidden" name="_token" value="{{ csrf_token() }}">
                <button type="submit" class="btn btn-primary">See status</button>
            </form>
        </div>

        @if(isset($map))
        @if($status=='Bumper To Bumper'))
        <div class="alert alert-danger" style='margin-left: 15%; margin-right: 15%;'>{{ $status }}</div>
        @endif
```

```
@if($status=='Clear'))
<div class="alert alert-success" style='margin-left: 15%; margin-right: 15%;'>{{ $status }}</div>
@endif
@if($status=='Moderate'))
<div class="alert alert-warning" style='margin-left: 15%; margin-right: 15%;'>{{ $status }}</div>
@endif
<br />
<?= $map ?>
@endif



<?php if (isset($name) && $name == -1) { ?>


    <div class="alert alert-warning" style='margin-left: 15%; margin-right: 15%;'>Sorry! traffic update for this location has not been updated.</div>
<?php } else { ?>


    <?php if ((($lat != '') && ($long != '')) || ($map != '')) {
        ?>
        <div id="map-canvas" style="width: 100%; height: 450px; margin-top: 20px; clear: both;"></div>
    <?php } else {
        ?>
        <div id="map-canvas" style="width: 100%; height: 450px; margin-top: 20px; clear: both;">
            <iframe
src="https://www.google.com/maps/embed?pb=!1m18!1m12!1m3!1d63821.14436821572 6!2d36.833406442644204!3d-1.2807770747829985!2m3!1f0!2f0!3f0!3m2!1i1024!2i768!4f13.1!3m3!1m2!1s0x182f1172d84d49a7%3A0xf7cf0254b297924c!2sNairobi%2C+Kenya!5e0!3m2!1sen!2s!4v1436097580401" width="100%" height="450" frameborder="0" style="border:0" allowfullscreen></iframe>
        </div>
    <?php } ?>
```

```
        <?php } ?>
    </div>
    <script src="public/js/bootstrap.min.js"></script>
    <div id="footer">
        <div class="container">
            <p class="text-muted">All rights reserved &copy; Ma3 route</p>
        </div>
    </div>
@include('_footer')
Traffic chart visualizer
<script type="text/javascript">
    $(function () {
        $('#container').highcharts({
            chart: {
                plotBackgroundColor: null,
                plotBorderWidth: null,
                plotShadow: false
            },
            title: {
                text: 'Traffic status proportionality'
            },
            tooltip: {
                pointFormat: '{series.name}: <b>{point.percentage:.1f}%</b>'
            },
            plotOptions: {
                pie: {
                    allowPointSelect: true,
                    cursor: 'pointer',
```

```
                dataLabels: {
                    enabled: true,
                    format: '<b>{point.name}</b>: {point.percentage:.1f} %',
                    style: {
                        color: (Highcharts.theme && Highcharts.theme.contrastTextColor) || 'black'
                    }
                }
            }
        },
        <?php
    $count = array_count_values($status_arr);
?>
        series: [{
            type: 'pie',
            name: 'Traffic Proportions',
            data: [
                ['Moderate', <?= isset($count['Moderate']) ?  $count['Moderate'] :  0 ?>],
                ['Bumber to bumber', <?= isset($count['Bumper To Bumper']) ?  $count['Bumper To
            Bumper'] :  0 ?>],
                ['Clear', <?= isset($count['Clear']) ?  $count['Clear'] :  0 ?>]
            ]
        }]
    });
});
</script>
```