# INTEGRATION OF TRAFFIC DATA FROM SOCIAL MEDIA AND PHYSICAL SENSORS FOR NEAR REAL TIME ROAD TRAFFIC ANALYSIS

UTSAV SONI
Mar, 2019

SUPERVISORS:
Mr. Kapil Oberai
Dr. Frank O. Ostermann

# INTEGRATION OF TRAFFIC DATA FROM SOCIAL MEDIA AND PHYSICAL SENSORS FOR NEAR REAL TIME ROAD TRAFFIC ANALYSIS

UTSAV SONI
Enschede, The Netherlands, March, 2019

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geoinformatics

SUPERVISORS:
Mr. Kapil Oberai
Dr. Frank O. Ostermann

THESIS ASSESSMENT BOARD:
Prof. Dr. Ir. A. Stein (Chair)
Prof. Dr. Raul Zurita-Milla (ITC Professor)
Dr. R. D. Garg (External Examiner, IIT, Roorkee)

DISCLAIMER

This document describes work undertaken as part of a programme of study at the Faculty of Geo-Information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

# ABSTRACT

Traffic congestion has become a serious problem in the present scenario. The economy of a country mostly relies on the transportation system of the country. The development of new transportation system require large amount of money and time as well. For a developing country like India, the best approach is to improve the efficiency of the existing system. This can be achieved by making use of the traffic related data from different sources and provide a real time information system to the people. The attainment of this information became possible due to the availability of the data from inductive loops installed at 5 location in Dehradun city. The inductive loop data was available in xml format, which consists of information like speed of the vehicle, vehicle length, road width, class, headway, date and time of acquisition etc. A model based on vehicle density and speed was utilised to estimate the level of traffic congestion (low, medium and high) from the inductive loop data. To enrich this information, data from another source was required. Earlier GPS data has been used for the same but its results were not that promising due to the poor accuracy of the GPS. So, this study attempts to integrate social media data with the inductive loop data to enrich the information and disseminate it in real time. Among all the social media platforms, the Twitter data is gaining popularity among the researcher community to gather the information of the events occurring in real time. The data was in the form of tweets which were extracted by using Twitter API. This data was classified into 7 different classes that are accident, event, congestion, weather, diversion, construction and other by using Naïve Bayes classification algorithm with the classification accuracy of 79.15%. Also, the sentiment analysis of tweets was done to understand the sentiment behind the tweets. These sentiments could be positive, neutral and negative. To develop a real time information system, the data obtained from both of these sources; inductive loops and Twitter data, was integrated by using the fuzzy rule method. As a result a web GIS based effective information system was developed for the users. Through this platform the users will be able to get the real time information regarding the cause and level of congestion and also the historical trend of traffic. Also, the transport engineers can make use of this information to improve the traffic condition and road network. This study concludes that the Twitter data can be used effectively to supplement the missing information of the inductive loop data. And it can be efficiently used with inductive loop data to provide the real time information to the transport engineers and general public through the web GIS based platform. This study shows how the freely available social media data can act as a valuable asset in the field of traffic monitoring.

Keywords: Inductive Loops; Congestion Model; Naïve Bayes; Data Integration; Intelligent Transport System; Traffic; Twitter

# ACKNOWLEDGEMENTS

I would like to take this opportunity to thank everyone involved in helping me complete this course successfully. The last 18 months spent at IIRS and ITC have been filled with immense learning and memorable experiences. It has required tremendous support from a lot of people, whom I wish to thank from the bottom of my heart.

First and foremost, I express my deepest gratitude to my IIRS and ITC supervisors, Mr Kapil Oberai and Dr Frank O. Ostermann, without whom completion of this exciting research project wouldn't have been possible. I have got to learn immensely from both of them, which I am sure will be very beneficial to me for the rest of my life.

Further, I thank IIRS, ITC and our program coordinators Dr Sameer Saran and Dr Valentyn Tolpekin who have extended endless support and guidance due to which we benefitted the most from our course. Dr Sameer Saran was always there to listen to our requests every time and boost our confidence. I extend my special thanks to him.

My due acknowledgements to Anurag Kulshreshta and Mr Aniruddha Deshmukh, who were a source of immense support throughout my research project phase. Without their support and guidance, the practical implementation of this project would have been much difficult. Many thanks to all the students and dear friends from MSc batch 2017-2019 for making the most hospitable and lovable atmosphere throughout the duration of the program. All these moments will be dearly missed. I would also acknowledge the support extended by Harikesh Singh, Akshay Rai, Anjali Madhu, Nivedhitha Ajithkumar, Anushree Badola, Debvrat Varshney and Abhisek Maiti for the social twitter experiment.

Finally, I thank my family and friends for being a constant source of support and encouragement. I will be in debt to all, always.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS

| Abbreviation | Full form |
|---|---|
| **AJAX** | Asynchronous JavaScript and XML |
| **API** | Application Program Interface |
| **CSS** | Cascading Style Sheets |
| **GIS** | Geographic Information System |
| **GPS** | Global Positioning System |
| **HMM** | Hidden Markov Model |
| **HTML** | Hypertext Markup Language |
| **ITS** | Intelligent Transport Systems |
| **JSON** | JavaScript Object Notation |
| **LIDAR** | Light Detection and Ranging |
| **LWR** | Lighthill-Whitham-Richards |
| **MFRI** | Mamdani Fuzzy Rule based Integration |
| **NB** | Naïve Bayes |
| **NBM** | Multinomial Naïve Bayes |
| **OSM** | Open Street Maps |
| **PCUPH** | Passenger Car Unit per Hour |
| **PHP** | Pre-Hypertext Processor |
| **POS** | Part Of Speech |
| **RADAR** | Radio Detection And Ranging |
| **SVM** | Support Vector Machine |
| **TF-IDF** | Term Frequency – Inverse Document Frequency |
| **XML** | eXtensive Markup Language |

# 1. INTRODUCTION

## 1.1. Background and Motivation

The urban cities of India have been facing the problem of traffic congestion at a rapid growth, which has its effect on the lifestyle, pollution, health, and fuel consumption. In the long run, traffic congestion has been identified as a barrier to social and economic growth (Delhi Economic Survey, 2015). It has now become a matter of great urgency to equip the cities in India with a transportation system which is efficient as well as environmentally sound to sustain the economic growth and improve the quality of life (Centre for Science and Environment, 1989). Traffic congestion studies have been started since 1925 by Charles Adler Jr. who designed the first Intelligent Transport System (ITS) which was made to govern the traffic movement speeds (Vinsel Lee, 2016).

The pattern of transport is different for different cities and is dependent on various factors like size, form, city structure, economic status, industrial & commercial growth and also its topography. The rapid increase in the transport vehicles with insufficient infrastructure and public services has been a growing issue in many cities. The current issues of navigation, congestion, traffic accident, pollution etc. are growing with the increase in transportation demand which has led to various studies contributing to the mitigation of these issues (Fazal, 2006). These studies have led to creation and updating of various policies as well as designing of efficient traffic monitoring and management systems or commonly known as Intelligent Transport Systems (ITS) (Davis, Joseph, Raina, & Jagannathan, 2017). Current ITS need quick response action on the emergencies, road facilities, flexible traffic operations for the travellers. Transport engineers require an improved system which can monitor and predict the status of the road networks and hence further help in taking decisions related to rerouting of mobility services, planning new transport infrastructure and ultimately result into measures which help reduce congestion and travel costs (Petalas, Ammari, Georgakis, & Nwagboso, 2017).

Traditional methods rely on data obtained from field-based sensors like single point loop detectors or clusters of detectors constrained between an intersection or a part of the corridor (Vlahogianni, Karlaftis, & Golias, 2014). Also, these traditional studies have mainly focussed on the traffic patterns at the intersections (Teodorovic, Lucic, Popovic, Kikuchi, & Stanic, 2001) or corridors (Lan, Sheu, & Huang, 2008; Schönhof & Helbing, 2007). There have been various different types of physical sensors utilized to generate traffic data. These sensors include Inductive loops, Pneumatic tubes, Magnetic sensors, Radio Detection and Ranging (RADAR) detectors, Light Detection and Ranging (LIDAR), Video cameras, Acoustic sensors, Bluetooth devices, and Global Positioning System (GPS) etc. However, the most preferable traffic sensor is inductive loop sensor which has proven to provide good sensitivity (Desai & Somani, 2014). Physical sensor data, although is a very reliable technology, but it is not cost effective (Leduc, 2008). Studies have been done to increase the efficiency of data collection by identifying the location to place the sensor (Eisenman, Fei, Zhou, & Mahmassani, 2006; Fei, Mahmassani, & Eisenman, 2007; Zhou & List, 2010). The existing studies on traffic movement patterns are specifically based on traffic flow (Michael J Cassidy & Bertini, 1999; Shen & Zhang, 2009; D. Zhang, Nagurney, & Wu, 2001), density (M. J. Cassidy & Mauch, 2001; Treiber & Kesting, 2010), speed (Banaei-Kashani, Shahabi, & Pan, 2011), etc. Further, studies have shown that combining direction, connectivity and locality of a road section can be used to determining traffic pattern with a high probability (Banaei-Kashani et al., 2011; Lv, Chen, Zhang, Duan, & Li, 2017). Although these might not be enough to explain traffic scenario in a huge road network with hundreds of intersections. Also it is impractical to provide sensor for each and every intersection with maintenance cost of an inductive loop detector ranging from $9,500 to $16,700 annually (Leduc, 2008; Lv et al., 2017).

The increasing demand of informative traffic knowledge like the purpose of the trip, better emergency response, etc. have made data mining and knowledge findings from multiple data sources a popular approach. In such a scenario, crowdsourcing data from various sources can be of a great use and it may be the only solution to a better traveller information system (Petalas et al., 2017). The sudden growth of available traffic data in the recent years have gained huge attention and has led to better management and control of ITS. Loop detector data, along with geosocial data, can be of great use in mining important information related to traffic operation and further provide insight of traffic congestions, accidents and help understand traffic patterns. The extraction, analysis, visualization and storage of the data puts focus on the two important views, syntactic and semantic. Syntactic view corresponds to the syntax while semantic view focuses on ways to generate meaning out of data by means of data analytics, manipulation, and visualisation. Social media are Web 2.0 based interactive Internet based applications. With the coming age of Social media, it has evolved as a reliable source of traffic information and also provides supplementary information for physical sensors (Kurkcu, Morgul, & Ozbay, 2015; D. Zhang et al., 2001). It gives opportunity to make use of every social media user as a social media sensor which provide real world information on the social media platforms. There have been many studies where road traffic information has been extracted from social media feeds (S. Zhang, Tang, Wang, & Wang, 2015).

In principle, it is advantageous to use multiple source data than using single source data (Hall & Llinas, 1997). Integration of features inherited from the tweet contents are expected to produce more informative results. Also such data can give information regarding traffic accidents which causes abnormality in the patterns of the traffic movement. So it sounds as a feasible method of monitoring traffic operations (Coifman et al., 1998; Oh, Oh, Ritchie, & Chang, 2005). There is an urgent need to design a systematic approach for traffic monitoring and analysis which takes into consideration both physical sensor data as well as social media data which covers spatial as well as temporal perspectives and provides near real time solutions to the current traffic problems.

## 1.2. Problem Statement

Data integration can be defined as the process of merging of data available from multiple data sources. This data could be available in structured as well as unstructured form and can be obtained from different sources like GPS, social media, inductive loops, magnetic sensors, video feed etc. Data integration tends to provide multidimensional information which has more significance than information from a single data source. Previous studies have been there for fusing social media and physical data sources for understanding traffic conditions but most of the studies were based on fusion of GPS data with respective geosocial data (Pan, Zheng, Wilkie, & Shahabi, 2013; S. Wang, Li, Stenneth, & Yu, 2016; S. Wang et al., 2017; Xu, Li, & Wen, 2018). Some of these studies provided real time analysis while some did analysis for a fixed interval. The drawbacks of using GPS based sensor is its vast amount of the errors in the dataset and poor accuracy of the results. These drawbacks are due to inefficiency in identification of location by the sensor, which caused false informations. A few studies also attempted to fuse social media data with fixed sensor data (Daly, Lecue, & Bicer, 2013; Giridhar et al., 2014; Lécué et al., 2014; Z. Zheng et al., 2018) but were carried out for a fixed duration and lack the near real time analytics aspect.

This research focuses on the gap of near real time provision of data integration of social media data with the physical traffic data which will help further the traffic engineers of the cities to understand and analyse the traffic situation.

## 1.3. Research Identification

This research aims to provide solution for integration of heterogeneous data available in the form of structured sensor data and unstructured traffic related social media data and address issues like traffic congestion estimation, traffic incident detection and historical trend analysis.

### 1.3.1. Research Objectives

The specific research objectives are:

1. To semantically and syntactically integrate the heterogeneous data available from traffic sensors and social media feeds.
2. To develop a prototype Web GIS application for provision of near real-time traffic situation monitoring systems for the transport engineers including the analysis of historical trends by using different data sources.

### 1.3.2. Research Questions

Related to the specific research objectives are the following research questions:

1. Research Objective 1:
    a. How do the various social media platforms compare with respect to their utility as traffic information source?
    b. How to detect spatial extent of non-geotagged traffic related social media feeds?
    c. How to integrate the qualitative data from social media feeds and quantitative data from traffic sensors?
2. Research Objective 2:
    a. Which is the most suitable congestion model for the available dataset?
    b. How to supplement the missing segments of sensor dataset with the social media feeds?
    c. How to handle and store the integrated data for historical trend analysis and for near real time?

### 1.3.3. Innovation aimed at

This study focuses on providing a near real time solution for integration of heterogeneous traffic data available in form of traffic flow and occupancy data available from the sensors and qualitative traffic related social media data

### 1.4. Thesis outline

This thesis comprises of seven chapters. **Chapter 1** introduces the motivation, problem statement, research objective, and research questions. **Chapter 2** provides the theoretical background of traffic data collection techniques, congestion models, geosocial data extraction and data integration. **Chapter 3** gives a detailed description of the study area, sensors and datasets used in this study. **Chapter 4** explains the methodology used to achieve the research objectives. **Chapter 5** describes and analyses the results achieved after executing the methods described in the previous chapter. **Chapter 6** concludes the research with answers to research questions and further recommendations.

# 2. LITERATURE REVIEW

Over the past decade, traffic data integration has been adopted increasingly due to it having many advantages. It has been observed that for accurate estimation and prediction of traffic based parameters available from various different sources, data integration is the one of the most suited approach (R. A. Anand, Vanajakshi, & Subramanian, 2011; Bachmann, 2011). For instance, a multi-data integration model using GPS and geosocial data was developed to estimate traffic congestion. This was done to eliminate the GPS errors which occur at time of traffic density estimation (Pan et al., 2013; S. Wang et al., 2017). While, Chu, Oh, & Recker (2005) used a Kalman filter model for integrating loop detector data and GPS data so as to predict travel time. Over time, different integration techniques like Kalman filters and others which makes use of Bayesian networks, artificial neural networks, or fuzzy logic rule based integration methods have gained acceptance in the field of ITS (Faouzi & Klein, 2016).

The experimental results from the previously stated studies gave promising results, although, none of them provided a near real time solution for integrating physical sensor data with data obtained from social media. The proposed traffic data integration technique attempts to address this aspect by providing a near real time solution to the problem.

This chapter has been subdivided into four major sections which provide the literature review on different road traffic data sources, social media and its applications in transportation, the various congestion models considered for the purpose of evaluating congestion with respect to various input parameters and finally a review on the different types of data integration algorithms used for the purpose of traffic congestion estimation.

## 2.1. Road Traffic Data Sources

### 2.1.1. GPS data

Making use of data mining models for GPS data using clustering and classification have been done in past (Kaklij, 2013). Thant Lwin & Thu Naing(2015) made use of Hidden Markov Model (HMM) to predict the road congestion using the GPS data collected from mobile phones on vehicles. There have been studies on traffic pattern analysis for segments of streets using GPS data (Necula, 2014, 2015).

### 2.1.2. Simulation data

Kim & Suh(2014) made use of multimodal simulation packages to interpret the variation between traffic flow inputs from standard sources. Metkari, Budhkar, & Maurya (2013) have made a simulation model for heterogeneous traffic data to predict traffic status. There also have been studies making use of spatiotemporal simulation for the purpose of traffic situation estimation (S. He, 2012).

### 2.1.3. Sensor data

Aslam, Lim, & Rus(2012) produced a congestion-aware routing system for traffic making use of physical sensor data. Urban traffic management systems make use of wireless sensors to estimate traffic congestion (Nellore & Hancke, 2016).

### 2.1.4. Probe data

X. Wang et al. (2015) presented a hidden Markov model making use of probe data to estimate urban vehicles. While, Hofleitner, Herring, Abbeel, & Bayen (2012) developed a real-time prediction of traffic dynamics using Bayesian network based on historical data available from probe vehicles.

## 2.2. Review On Social Media

Social media are Web 2.0 based interactive internet based applications. It acts as a collection of various different communication channels which runs on community based inputs and allows individuals, companies, government and organizations to create and share their respective interests (Asur & Huberman, 2010). According to (Kaplan & Haenlein, 2010) it comprises of following different services:

- Provision of adding user-generated content like textual posts and media files.
- Provision to create specific profiles for websites, apps, businesses.
- Provision of connecting with groups of people of similar background and interests.

The social media has been existing since long time in form of traditional news, media or website and has enhanced a lot with the transformation of various communication technologies. The technical evolution has been responsible for moving of Social Media applications from desktop PCs to mobile gadgets. The evolution is responsible for creation of bidirectional information sharing platforms which have led to various popular social media websites such as Facebook, Instagram, Google+, Twitter, Pinterest, LinkedIn etc.

Table 2.1 showcases the statistics of monthly active users of the above mentioned social media website (Statista, 2018).

Table 2.1: Social Media statistics (Statista, 2018)

| Social Media Website | Monthly Active Users (in millions) | Social Media Release Year |
|---|---|---|
| Facebook | 2320 (Dec, 2018) | 2004 |
| Instagram | 1000 (Jun, 2018) | 2010 |
| Google+ | 111 (Apr, 2013) | 2011 |
| Twitter | 321 (Dec, 2018) | 2006 |
| Pinterest | 250 (Sep, 2018) | 2010 |
| LinkedIn | 260 (Jun, 2018) | 2002 |

Social media technologies are available in many different forms like blogs, enterprise networks, photo sharing, forums, social networks, video sharing etc. (Aichner & Jacob, 2015). Each social media platform has its own specialization and focus like microblogging, social network, photo sharing, video sharing, career networks etc.

### 2.2.1. Application of social media in transportation

Despite availability of various different types of social media, there are various common features in social media which make it apt for usage in transportation industry. Social media can reflect the different ongoing trends provided by massive crowd of people which makes it possible to perform studies like traffic accident studies, congestion estimation, and social event detection, etc. The benefits of making use of social media to carry out these studies are:

- Social media are based on mainly online data which can be extracted by means of Application Program Interface (API). One of the most popular API is from Twitter Inc.
- Since social media acts as the reflection of social connections among people, almost all existing activity feed can be found in social media.

Social media comes along not with not just benefits but also with some problems, some of these problems are:

- Noise: social media requires filtering and cleaning of the meaningless samples in the data.
- Unstructured content: unlike the traditional traffic data, the data available from social media are different; available in form of words, pictures etc. which further needs to be categorized.
- Untrustworthy: social media have abundance of fake and untrustworthy data, hence cross validation is necessary.
- Location parameters: for transportation studies, location plays an important role although the location data available from social media is either available in textual form which can be vague or in form of GPS data which may consist GPS errors.

To counter these challenges, various technologies like natural language processing, sentiment analysis etc. have to be carried out. Table 2.2 provides an overview of different transportation applications studies using social media data and the type of social media used.

Table 2.2: Social Media for Transportation Application

| Transportation application | Social media used | Author |
|---|---|---|
| Transportation planning | Facebook, Twitter, Flickr etc. | (Camay, Brown, & Makoid, 2012) |
| | Facebook, Twitter, YouTube | (Stambaugh, 2013) |
| | Twitter | (Chan & Schofer, 2014) |
| Travel information retrieval | Google Search, Twitter | (Xiang & Gretzel, 2010) |
| | Twitter | (Sasaki, Nagano, Ueno, & Cho, 2012) |
| | Twitter | (J. H. Lee, Gao, & Goulias, 2015) |
| Traffic incident detection | Twitter | (Schulz, Ristoski, & Paulheim, 2013) |
| | Twitter | (S. Zhang et al., 2015) |

It can be noted that most of the social media studies utilize Twitter data as a source. This is because Twitter data provides easy access to spatial as well as temporal information and also exact coordinates in case of geo-tagged tweets. The large amount of collected data can be easily accessed by means of Twitter API (Twitter, 2011). For case of other social media platforms, 3rd party methods have to be used so as to extract the data and with certain selected search criterion only. Thus it is not possible to collect large quantities of data from other social media platforms for research purposes.

### 2.2.2. Traffic activity detection based on Twitter data

#### 2.2.2.1. Traffic congestion detection

The current state-of-art studies consider the impact of human activities on traffic congestion. This leads to diversity in answers to correlation between the two aspects. Over the past decades, studies have made use of social media to validate major events like natural disasters (Thianniwet, Phosaard, & Pattara-Atikom, 2010), political events (Kiilu, Okeyo, Rimiru, & Ogada, 2018), etc. The problem arises to find whether it can detect traffic congestion and to identify the correlation between tweets and traffic congestion. Usually, tweets related to a specific event will be frequently available and it can be used to find direct correlation between the tweets and the events (Amin Elsafoury, 2013). However, traffic related tweets may be different from that of other events because not many report the same traffic situation of same spatial and temporal context (Giridhar et al., 2014). In this study, the focus is on the tweets that involve a wide variety of traffic related activities and find correlation between the tweets and traffic congestion

#### 2.2.2.2. Traffic incident detection

Traditional studies have been focussing mainly on monitoring traffic flow changes to detect traffic events due to the fact that major events have higher impact on traffic flow disruptions. Some methods compared temporal & spatial features to detect anomalies in traffic flows which may indicate possibility of a traffic incident where the conditions are different from normal conditions. Despite the attempts by above studies, traditional methods using only traffic data to identify traffic incident was a challenging task. Most of the previous studies relying on just field data to detect traffic incident were working with the assumption that data is reliable. But there can be the chances of sensor failure, due to which false information can be obtained regarding traffic situation. Hence, the abnormalities in the traffic patterns observed by field data or sensor data is not enough to justify the traffic incident.

Hence, to overcome these challenges, instead of relying on physical sensors, Twitter, the microblogging platform has been recently used by many studies to gather crowdsourced information for incident detection. Twitter provides an online platform where a user can create, utilize, promote or share content for certain communities (Chan & Schofer, 2014). Hence, each tweet acts as a social sensor for retrieval of wide range of information from large groups of people in a timely manner. Moreover, with the increased usage of mobile devices for posting tweets, availability of corresponding location and time makes Twitter a great source for detection of traffic incidents (White, Thompson, Turner, Dougherty, & Schmidt, 2011).

### 2.2.3. Twitter traffic data sentiment analysis

From the time Twitter has launched in 2006, application of sentiment analysis has been implemented in various domains. The data available from Twitter could be from various different road traffic users who are expressing their opinions on various traffic situations ranging from traffic jam, accidents, diversions, road closure etc. The main question which arises is how to determine the traffic state from the textual information provided in a Tweet. This is where sentiment analysis comes into role. Sentiment analysis can be defined as an in-depth process of computational extraction of opinions from the text and identify if the emotional attitude towards the topic is positive, negative or neutral (Pak & Paroubek, 2010). Kumar & Sebastian (2012) presented the method of sentiment analysis by making use of dictionary based methods to estimate the orientation of keywords present in the tweets.

J. He, Shen, Divakaruni, Wynter, & Lawrence (2013) considered applying sentiment analysis for estimation of traffic state prediction, and then examined correlation between traffic volume and tweet count. Abidin, Kolberg, & Hussain (2015) made use of Kalman filter model for travel time estimation using traffic data from Twitter API as input. Studies have also provided real-time traffic flow estimation by sentiment analysis of social media data (Grosenick, 2012).

## 2.3. Congestion Models

Traffic data available from various sources is usually quantitative in nature. In order to perform the congestion study, there is a need to transform it into qualitative data. This is where congestion modelling comes into role. It provides the information of severity of the traffic in qualitative manner i.e. low, high and medium etc. Congestion in roads is directly related to the travel time, delay, speed, density of vehicles in a road, traffic volume etc. There have been multiple studies which have made use of various congestion models to categorize the traffic data to different congestion levels.

Y. Wang & L. Nihan (2005) proposed an approach of using results from single loop detector and used a model based on estimated speeds of Freeway to produce traffic congestion information. One of the study presented a model and algorithm for estimating traffic congestion. They used traffic data as congestion factor to determine level of congestion making use of Jamitons and Lighthill-Whitham-Richards (LWR) models using information about speed and traffic flow density (Duan, Liu, & Sun, 2009). Fusco, Colombaroni, & Sardo (2012) made use of Artificial Neural Networks, automatic incident detection and road traffic network model to predict the traffic congestion level. In a study conducted by Henry & Koshy (2016), they have developed congestion model by considering time travel index that is a ratio of actual time taken by a vehicle to the time taken by it during free flow. They have developed this model for a road stretch lies between Kumaranalloor and Gandhi Nagar in Kottayam, Kerala India. Another study has been done by Thianniwet, Phosaard, & Pattara-Atikom (2010), they have considered instantaneous speed with the average speed of the vehicle. They have classified congestion as light, heavy and jam. Although, speed alone cannot be a good indicator for congestion. Because sometimes there might be a case that even at very low congestion, the vehicle will move slowly because driver may not be in hurry. So inclusion of multiple parameters is a necessity for modelling congestion. In a study done by Maitra, Sikdar, & Dhingra (1999), they have built an empirical model for congestion considering the speed of the vehicle and the density of the vehicles in the road at a particular time. Although it is an old model, but till now people are using this model for congestion modelling because it is the most suitable model particularly for the Indian roads (Patel & Gundaliya, 2016).

## 2.4. Data Integration

Data integration in context of this study involves integration of multi-sensor data. It consists of combining different structured with unstructured data. Data integration has the benefit of improved estimation of observed measurements. Also data integration tends to reduce the uncertainty in the data and makes it more dependable (Elmenreich, 2002). For accurate and reliable traffic flow congestion estimation, data integration is often done using powerful algorithms like Bayesian network, Kalman filter etc. One of the study applied Kalman filter to integration spatial & location based data using a time series regression model (A. Anand, Ramadurai, & Vanajakshi, 2014). While, Sun, Guo, Liu, Feng, & Hu (2009) identified integration of inductive loop data and GPS data at different levels namely, data level, feature level and decision level making use of bayesian and entropy based methods for traffic data integration.

### 2.4.1. Data integration algorithms for estimating traffic congestion

Table 2.3 provides a list of different data integration algorithms used in the past for estimating traffic congestion.

Table 2.3: Data integration algorithms for estimating traffic congestion

| Algorithm | Author |
|---|---|
| Kalman Filter | (A. Anand et al., 2014)<br>(Jiann-Shiou Yang, 2005) |
| Extended Kalman Filter | (S.-S. Kim & Kang, 2007)<br>(Guo, Xia, & Smith, 2009) |
| Bayesian Network | (Sun et al., 2009)<br>(Pamuła & Król, 2016)<br>(W. Zheng, Lee, & Shi, 2006) |
| MFRI | (Awan & Awais, 2011)<br>(K. Runyoro & Ko, 2013)<br>(T. O. Adetiloye, 2018) |

# 3. STUDY AREA , DATASETS, HARDWARE & SOFTWARE TOOLS

## 3.1. Study Area

Dehradun is the capital of Uttarakhand, a state in northern part of India. It is located on the foothills of the Himalayas at an altitude of 640 metres, between the rivers Ganga and Yamuna. Being a recent capital, Dehradun has witnessed very high and sudden growth in terms of urban population and traffic due to which the current transportation infrastructure faces heavy congestion problems on daily basis. Two major highways (NH 72 and NH 58) pass through this city and most used route to reach the famous hill station Mussoorie passes through the central part of the city. This makes it a good choice spot for conducting this study.

## 3.2. Required Tools

This section list down all the datasets, hardware and software used for execution of this research project.

### 3.2.1. Hardware Used

Within the city of Dehradun, there are 5 inductive loop based sensors based M680 vehicle counter and classifier installed at major road junctions as highlighted in Figure 3.2.

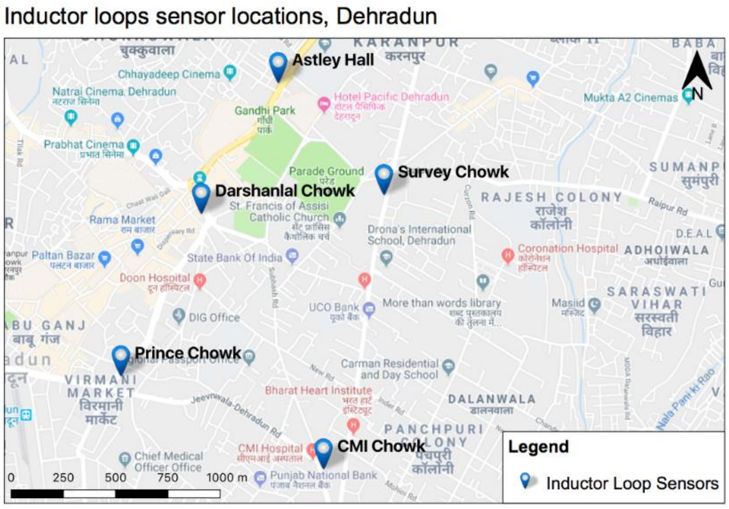


Figure 3.2: Sensor Locations

Figure 3.1: M680 Traffic counter

Figure 3.3 shows the inductive loop installation at different locations in Dehradun. The inductive loops are installed underground while the counter which is responsible for accessing and storing information is installed on poles beside the loops.

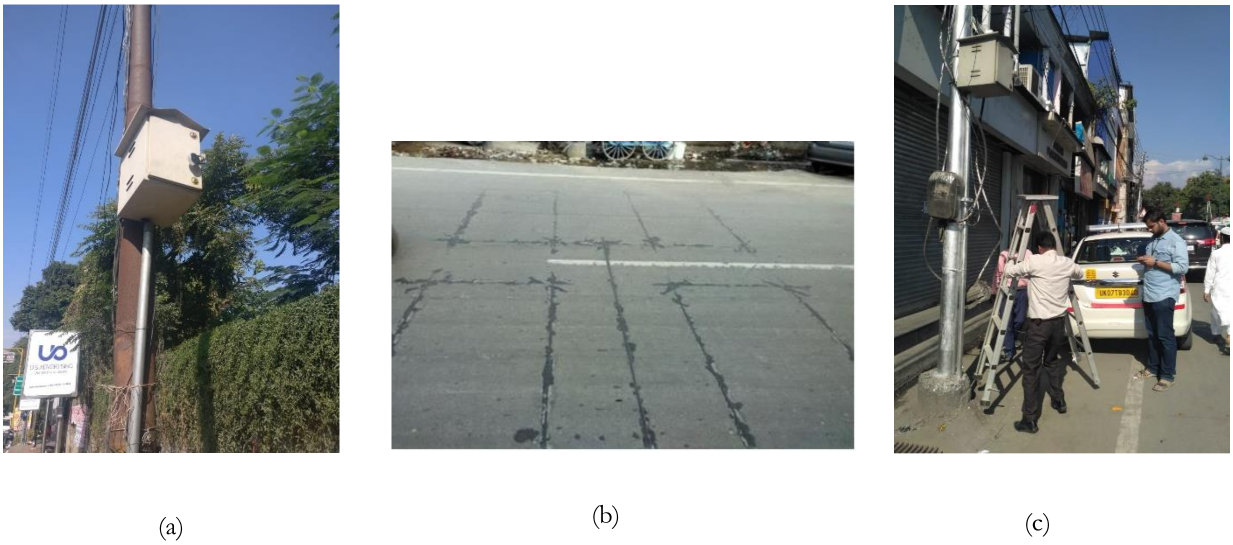


Figure 3.3: Photographs showing sensor installation in Dehradun city where (a) shows traffic counter installed on the pole next to the bus stand on CMI Chowk, (b) shows the installed loops at Astley hall and (c) shows the process of maintenance of the sensors on the field.

### 3.2.2. Software technologies used:

This section provides a list of the technologies utilized for completion of this project. Table 3.1 Table 3.2

Table 3.1: Web technologies used to develop web GIS based dashboard application

| Technologies | Functions |
|---|---|
| **HTML**<br>**CSS** | Used for creating the basic user interface structure of the application. |
| **JavaScript** | Used for providing interactivity of the elements |
| • **JQuery** | JavaScript framework for optimised scripting |
| • **Leaflet JS** | JavaScript framework for working with interactive maps |
| o **Leaflet-Heat.JS** | Leaflet plugin for creating heat maps |
| o **Leaflet-GoogleMutant.JS** | Leaflet plugin for integration with Google Maps API |
| • **AJAX** | Used for making asynchronous calls to the server side for executing database related queries in run time. |
| **PHP** | Server side language for accessing the database |
| **MongoDB** | NoSQL based data storage method |

Table 3.2: List of python libraries used for this project

| Library | Package | Functions | Usage | Source |
|---|---|---|---|---|
| afinn | - | Afinn | To evaluate sentiment score for tweets | (Nielsen, 2011) |
| collections | - | defaultdict | Dictionary for Parts of Speech (POS) tagging of tweets | |
| json | - | | To convert | (JSON, 2019) |
| math | - | | To use basic mathematical functions for the congestion model | |
| nltk | corpus | stopwords, | Stopwords for filtering from the text | (Bird, Edward, & Ewan, 2009) |
| | | wordnet | Wordnet | |
| | stem | WordNetLemmatizer | For lemmatization | |
| | tokenize | word_tokenize | Convert statement to token of words | |
| | - | pos_tag | To assign the type of speech to each word | |
| numpy | - | | To deal with arrays | (Numpy, 2018) |
| os.path | - | | To read directories | |
| pandas | - | | To work with data frames | (Pandas, 2019) |
| pymongo | - | | To connect python to MongoDB | (MongoDB, 2008) |
| pytz | - | | For time zone calculations | |
| re | - | | For dealing with regular expressions | |
| sklearn | preprocessing | LabelEncoder | To assign numerical codes to labels | (Pedregosa et al., 2011) |
| | feature_extraction.text | TfidfVectorizer | For transforming textual words to vectors | |
| | - | model_selection | For splitting testing & training data | |
| | - | naive_bayes | For textual classification | |
| | metrics | accuracy_score,<br>cohen_kappa_score | For accuracy assessment<br>For evaluating kappa score | |
| tweepy | - | - | For accessing Twitter streaming API | |

### 3.2.3. Dataset Used

#### 3.2.3.1. Inductive Loop Sensor

The data of inductive loops which are installed across Dehradun is collected. This data is available is available in form of near real time stream as well as historical data. The historical data is available in space delimited textual format as shown in Table 3.4 and the near real time data is available in XML format as shown in Table 3.3.

Table 3.3: Sensor near real time streaming data

```
<vehicle version="1.2">
  <time>2017-12-05T23:10:00.100</time>
  <lane>0</lane>
  <subSite>1</subSite>
  <speed>36</speed>
  <length>3</length>
  <headway>23.9</headway>
  <gap>23.7</gap>
  <direction>false</direction>
  <class scheme="eur6">1</class>
  <chassisHeightCode>4</chassisHeightCode>
  <occupancyTime>300</occupancyTime>
</vehicle>
```

Table 3.4: Sensor historical data

| HEAD | DDMMYY | HHMM | SS | HH | RESCOD | L | D | HEAD | GAP | SPD | LENTH | CS | CH |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 000001 | 040717 | 0000 | 01 | 50 | 000000 | 3 | 2 | 1.5 | 1.5 | 28 | 544 | 4 | H |
| 000003 | 040717 | 0000 | 16 | 10 | 000000 | 2 | 1 | 16.1 | 16.1 | 36 | 100 | 1 | H |
| 000004 | 040717 | 0000 | 28 | 70 | 000000 | 2 | 1 | 12.6 | 12.5 | 36 | 355 | 2 | L |

#### 3.2.3.2. Twitter Stream

Twitter Streaming API provides a real time stream of the tweet data on the basis of passed queries. In this research, the search query will be used where the keywords are selected and a streaming Real time twitter data extracted using Twitter developer tools based on the search query which is streamed in JavaScript Object Notation (JSON) format.

# 4. METHODOLOGY

This chapter explains about the methodology utilized for the implementation of this research project. The general methodology which has been followed has been displayed in Figure 4.1. It consists of four major sections, i.e. Twitter data analysis, Sensor data analysis, Data integration and Data visualization which are further discussed. This chapter describes the study area and the datasets utilized for this research and also gives a brief about the software and hardware used and then explains in details all the steps involved in the execution of this research project.

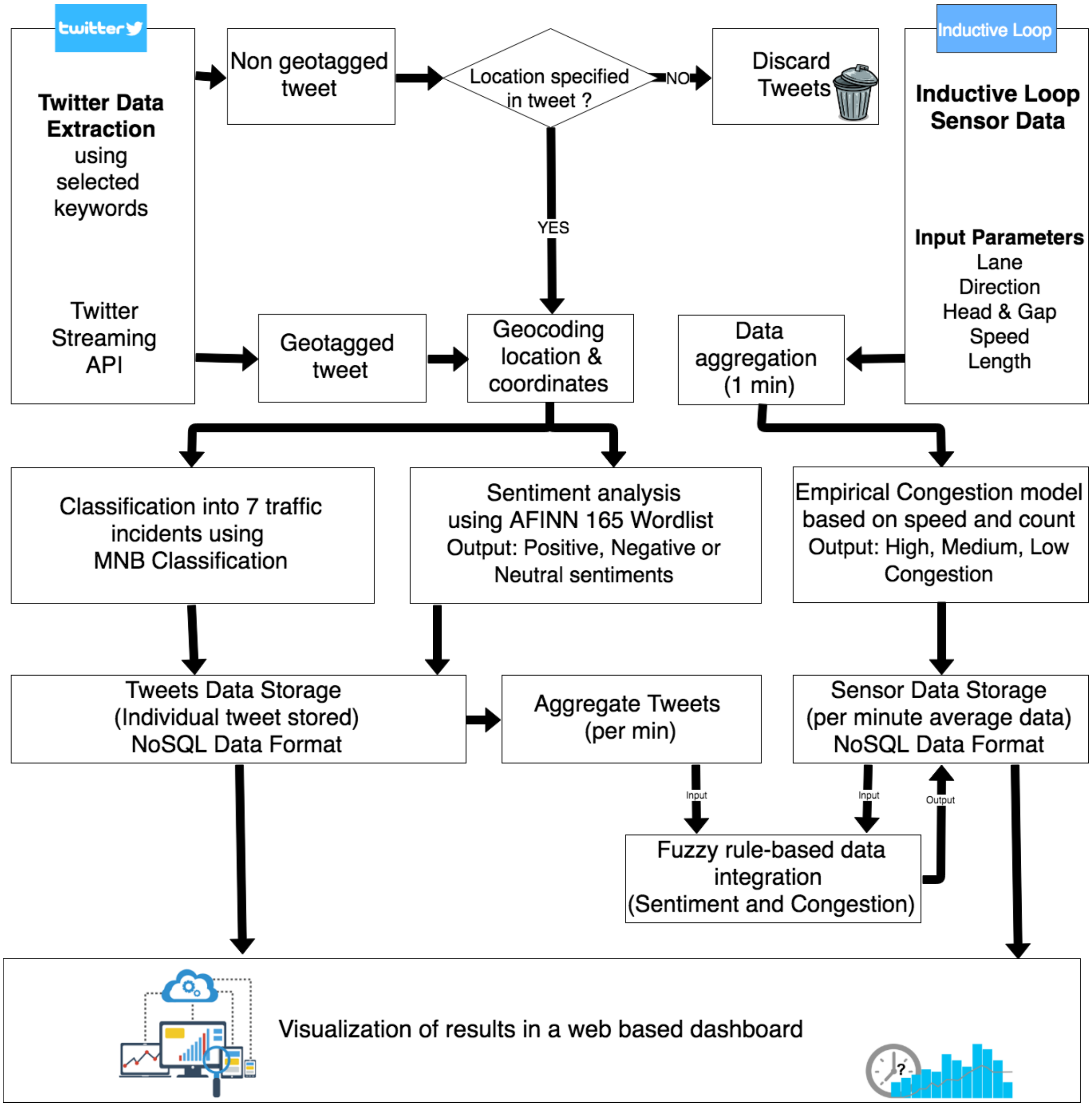
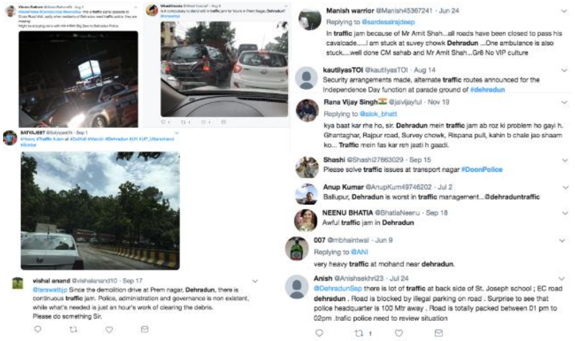

twitter
Twitter Data Extraction using selected keywords
Twitter Streaming API
Non geotagged tweet
Location specified in tweet ?
NO
Discard Tweets
YES
Inductive Loop
Inductive Loop Sensor Data
Input Parameters
Lane
Direction
Head & Gap
Speed
Length
Geotagged tweet
Geocoding location & coordinates
Data aggregation (1 min)
Classification into 7 traffic incidents using MNB Classification
Sentiment analysis using AFINN 165 Wordlist
Output: Positive, Negative or Neutral sentiments
Empirical Congestion model based on speed and count
Output: High, Medium, Low Congestion
Tweets Data Storage (Individual tweet stored) NoSQL Data Format
Aggregate Tweets (per min)
Sensor Data Storage (per minute average data) NoSQL Data Format
Input
Input
Output
Fuzzy rule-based data integration (Sentiment and Congestion)
Visualization of results in a web based dashboard

Figure 4.1: General Methodology

## 4.1. Twitter Data Analysis

Before beginning the process of extraction of tweets, it is a necessary step to analyze the availability of the data and set parameters for the search criteria that have to be considered for the extraction. For this purpose, traffic related tweets for Dehradun city were manually searched. It resulted in tweets available from both the general public as well as tweets available from Senior Superintendent of Police (SSP Dehradun, the traffic managing authority of Dehradun city) as shown in Figure 4.2 & Figure 4.3.

From all the collected tweets, it was observed that there were many tweets which were written in Hindi. It was also observed that people make use of different spellings and sometimes short forms for highlighting different keywords, like Dehradun, Dehradoon, and Ddoon etc. Although it is not possible to consider all the variants, so the ones with maximum results were considered. The common terms which were found in the traffic related tweets were used to create a dictionary of keywords which is further discussed in section 4.1.1. After deciding over the initial search parameters, the methodology shown in Figure 4.4 is used so as to extract, process and store Twitter data.



vishal anand @vishalanand10 · Sep 17
@tanswatbjp Since the demolition drive at Prem nagar, Dehradun, there is continuous traffic jam. Police, administration and governance is non existent, while what's needed is just an hour's work of clearing the debris.
Please do something Sir.

Manish warrior @Manish45367241 · Jun 24
Replying to @sardesairajdeep
In traffic jam because of Mr Amit Shah...all roads have been closed to pass his cavalcade......I am stuck at suvey chowk Dehradun ...One ambulance is also stuck....well done CM sahab and Mr Amit Shah...Gr8 No VIP culture

kautilyasTOI @kautilyasTOI · Aug 14
Security arrangements made, alternate traffic routes announced for the Independence Day function at parade ground of #dehradun

Rana Vijay Singh @jaivijayful · Nov 19
Replying to @alok_bhatt
kya baat kar rhe ho, sir. Dehradun mein traffic jam ab roz ki problem ho gayi h. Ghantaghar, Rajpur road, Survey chowk, Rispana pull, kahin b chale jao shaam ko... Traffic mein fas kar reh jaati h gaadi.

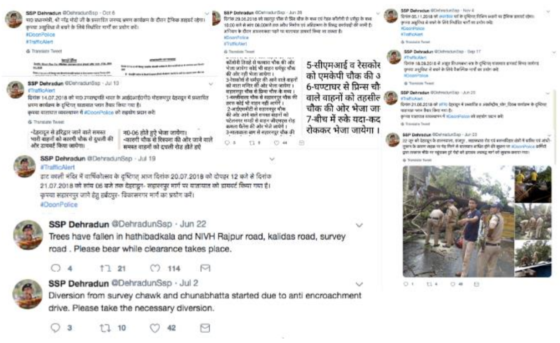
Shashi @Shashi27663029 · Sep 15
Please solve traffic issues at transport nagar #DoonPolice

Anup Kumar @AnupKum49746202 · Jul 2
Ballupur, Dehradun is worst in traffic management...@dehraduntraffic

NEENU BHATIA @BhatiaNeenu · Sep 18
Awful traffic jam in Dehradun

007 @mbhaintwal · Jun 9
Replying to @ANI
very heavy traffic at mohand near dehradun.

Anish @Anishsekhri23 · Jul 24
@DehradunSsp there is lot of traffic at back side of St. Joseph school ; EC road dehradun . Road is blocked by illegal parking on road . Surprise to see that police headquarter is 100 Mtr away . Road is totally packed between 01 pm to 02pm .trafic police need to review situation



SSP Dehradun @DehradunSsp · Jul 13
#TrafficAlert
दिनांक 14.07.2018 को मा0 उपराष्ट्रपति भारत के आई0आई0पी0 मोहकमपुर देहरादून में प्रस्तावित भ्रमण कार्यक्रम के दृष्टिगत् यातायात प्लान तैयार किया गया है।
कृपया यातायात व्यवस्थाओं में #DoonPolice को सहयोग प्रदान करें

5-सीएमआई व रेसकोर
को एमकेपी चौक की ओ
6-घण्टाघर से प्रिन्स चौ
वाले वाहनों को तहसील
चौक की ओर भेजा जा
7-बीच में रुके यदा-कद
रोककर भेजा जायेगा ।

SSP Dehradun @DehradunSsp · Jul 19
#TrafficAlert
डाट काली मंदिर में वार्षिकोत्सव के दृष्टिगत् आज दिनांक 20.07.2018 को दोपहर 12 बजे से दिनांक 21.07.2018 को सांय 06 बजे तक देहरादून- सहारनपुर मार्ग पर यातायात को डायवर्ट किया गया है।
कृपया सहारनपुर जाने हेतु हर्बटपुर- विकासनगर मार्ग का प्रयोग करें।
#DoonPolice

SSP Dehradun @DehradunSsp · Jun 22
Trees have fallen in hathibadkala and NIVH Rajpur road, kalidas road, survey road . Please bear while clearance takes place.
4 21 114

SSP Dehradun @DehradunSsp · Jul 2
Diversion from survey chawk and chunabhatta started due to anti encroachment drive. Please take the necessary diversion.
3 10 42

Figure 4.2: Tweets from general public related to Dehradun city

Figure 4.3: Tweets SSP Dehradun

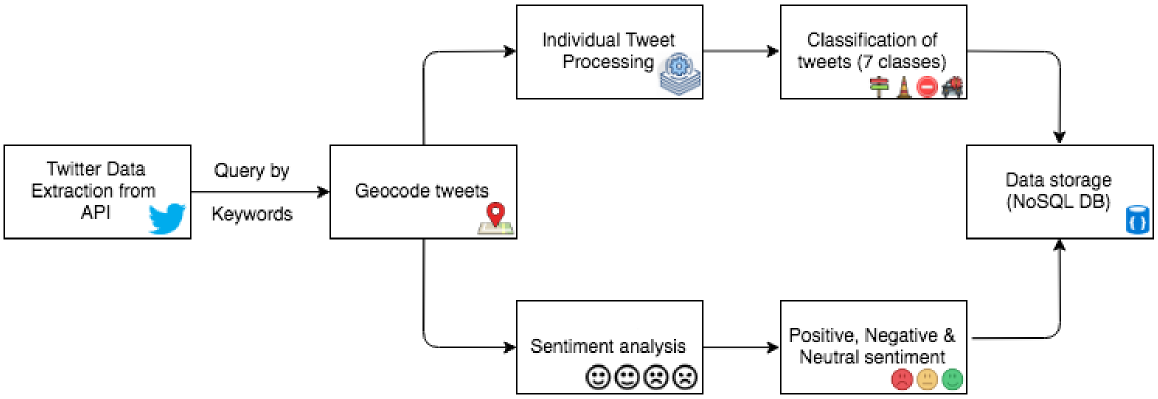

Individual Tweet Processing
Classification of tweets (7 classes)
Twitter Data Extraction from API
Query by Keywords
Geocode tweets
Data storage (NoSQL DB)
Sentiment analysis
Positive, Negative & Neutral sentiment

Figure 4.4: Twitter Data Processing Methodology

### 4.1.1. Twitter Data Extraction & Pre-processing

An Application Programming Interface (API) is a tool that provides easy interaction between computer programs and web services. Various web services have provided APIs so as to provide access to their data and services. The famous social media giant, Twitter also makes provision of its data and services by means of its Streaming API. Twitter has also made provision for developers to access its data and services by means of Streaming and REST API (Twitter, 2011). Streaming API provides a real time stream of data while REST API provides data on the basis of selective search terms but not in real time. This research makes use of Twitter's Streaming API to collect live tweets and their attributes on the basis of certain search based keywords which are relevant to the scope of the study. The keywords that are utilized for this purpose have been stated in Table 4.1. These keywords are selected after manually going through tweets available on Twitter relating to the traffic situation in Dehradun. On sending the request with the query parameters, streams returns a stream of response in JSON format. Further pre-processing is done on the data collected from the stream.

Table 4.1: Keywords pairs used as query to extract tweets from Twitter Streaming API

| | | | | |
|---|---|---|---|---|
| देहरादून traffic | देहरादून trafic | देहरादून यातायात | देहरादून जाम | देहरादून jam |
| dehradun traffic | dehradun trafic | dehradun यातायात | dehradun जाम | dehradun jam |
| dehradoon traffic | dehradoon trafic | dehradoon यातायात | dehradoon जाम | dehradoon jam |
| doonpolice traffic | doonpolice trafic | doonpolice यातायात | doonpolice जाम | doonpolice jam |
| dehradunssp traffic | dehradunssp trafic | dehradunssp यातायात | dehradunssp जाम | dehradunssp jam |

Twitter streaming API gives out a lot of attributes as response. Pre-processing of the data is required so as to filter out the non-relevant data from the response and keep just the relevant data. Parameters considered as include: *text, screen_name, id, created_at*, and *coordinates* which are explained in Table 4.2.

Table 4.2: Attributes extracted from Twitter Stream with explanation

| **Attribute** | **Explanation** |
|---|---|
| text | This is the content of the tweet. |
| screen_name | This is the username of the user who posted the tweet |
| id | This is the unique ID of the tweet |
| created_at | This is the date & time of the creation of the tweet in the form of:<br>(weekday month date time (in 24 hr format) + time zone year )<br>Example: (Sun Feb 03 23:48:36 +0000 2019) |
| coordinates | (Latitude, longitude) of the user who posted the tweet. This is available only if the user had enabled the option of sharing accurate location before tweeting. |

### 4.1.2. Geocoding Tweets

Geocoding is the process of extraction of location data from a textual phrase in form of its coordinates (latitude, longitude). While posting a tweet, one has to manually select option of sharing precise location before every tweet so that the exact latitude and longitude gets registered along with the tweet. This added effort makes people skip the step and hence there is scarce availability of geotagged tweets. This generates a need to find an alternate solution to identify location of the tweet to avoid data loss. This is where geocoding comes into role. There are various different geocoding libraries services available. In this study, common Hex geocoding API provided by Google Maps services has used. The text obtained from the tweet is passed upon as a query to the API and a response consisting of latitude, longitude and identified address is received which is stored. For the case of geotagged tweets as well, the retrieved latitude and longitude are passed as query to the API to reverse geocode the address of the tweet and is stored. Finally.

### 4.1.3. Sentiment analysis

Tweets are textual phrases which are limited to 140 characters length. It is a necessary task to find the sentiment involved with the tweet so as to understand the emotion with which the user has posted the tweet. This helps to understand the criticality of the situation which is expressed by the means of tweet. Since this research does not focus on understanding the emotion expressed by a user, hence for the purpose of sentiment analysis, AFINN, a common sentiment analysis library is used (Nielsen, 2011). AFINN library of sentiment analysis makes use of a wordlist named AFINN-165 which has list of words along with a score distributed on a scale of -5 to 5. For this research, the output of the sentiment analysis is categorized into 3 classes as shown in Table 4.3. These 3 classes are further utilized to be fed as an input for the integration model which is discussed in 4.3

Table 4.3: Sentiment score classification

| Cumulative Sentiment Score | Sentiment Class |
|---|---|
| Score > 0 | positive |
| Score = 0 | neutral |
| Score < 0 | negative |

### 4.1.4. Traffic incident classification

Sentiment analysis of tweets as discussed in section 4.1.3, only provides information related to the sentiment associated with the tweet. It is unable to cover the reason behind that sentiment or in this case, type of incident highlighted within the tweet. Hence, to extract information related to the incident it is necessary to classify the tweets into different classes. For this research, seven different classes have been considered which define the scenario explained within the tweet in form of a traffic incident as shown in Table 4.4.

| Incident Type |
|---|
| Accident |
| Congestion |
| Construction |
| Diversion |
| Event |
| Weather |
| Other |

Table 4.4: Traffic incident classes

For the purpose of classification, there is a need of selection of appropriate classifier. In area of natural language processing, Naïve Bayes (NB) algorithm has gained importance in the scientific community. NB classifier is a simple classifier which works on the Bayes theorem. It is a very popular classification technique and have outperformed the Random forest, Support vector machine (SVM), and Supervised linear regression (Li, Caragea, Caragea, & Herndon, 2016). In a study done for hate speech recognition from the Twitter data, NB classifier has shown good results with an accuracy of 67.47% (Kiilu et al., 2018). It can model only the presence and absence of words while in case where data can convert into counts easily like in case of words count in the text, Naïve Bayes Multinomial (NBM) classifier can be used. It is a specialized version of NB classifier and can be effectively used in natural language processing (Mccallum & Nigam, 1998). In a study done by K. Lee et al. (2011), they performed text based classification of the tweets and compared NBM, NB and SVM on the basis of accuracy. In this study NBM has outperformed the other two classifier by attaining an accuracy of 65.36%. Due to its good performance in case of classification of words, NBM is selected as the classifier for the classification of words. NBM classifier works on the principle of the Bayes theorem. It is given by:

$$P(c|d) \propto P(c) \prod_{1\leq k\leq n_d} P(t_k|c) \qquad \text{(Equation 1)}$$

Where,
$c$ is the set of classes used in the classification
d is the tweet related to traffic
$t_k$ is the tokens/ words of a tweet
$P(c|d)$ is conditional probability of class c given d
$P(c)$ is prior probability of class c
$P(t_k|c)$ is conditional probability of $t_k$ given class c

Machine learning algorithms do not take text as direct input for classification, hence the text needs to be pre-processed and converted into numerical forms. The pre-processing steps of tweet text followed in this research are explained in Figure 4.5.

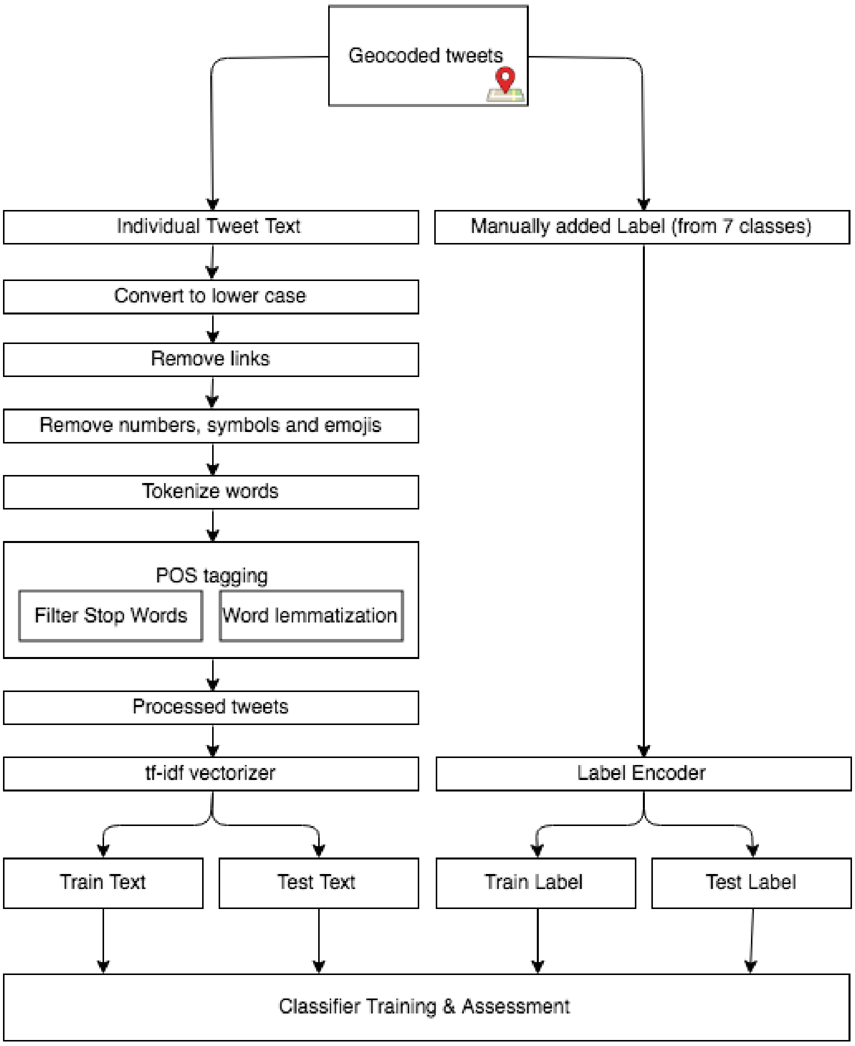


Figure 4.5: Text Pre-processing for classification. Note: POS is parts of speech and tf-idf is Term frequency-inverse document frequency.

In this study, scikit-learn library has been used for performing NBM classification. For the training of classifier, 1885 tweets have been extracted using the extraction method specified in section 4.1.1 have been used which are then been labelled manually.

### 4.1.5. Twitter Data storage

Once the Twitter data has gone through the process of filtering, geocoding, sentiment analysis and traffic incident classification, the final outcome is ready to be stored. The outcome is structured in dictionary format and then stored in a collection inside MongoDB.

## 4.2. Sensor Data Analysis

Vehicle by vehicle data is extracted from the inductive loop sensor. This extracted data is available in 2 different formats as explained in section 3.2.2. For sensor data analysis, real time filtering python script is used, which separates the collected vehicle data according to direction of movement and also removes the records where the abnormalities occur. Along with this, the script also identifies the type of vehicle on the basis of its length as shown in Table 4.5.

Table 4.5: Vehicle Classes

| Label | Class | Length (cm) |
|---|---|---|
| L | Light | length < 250 |
| M | Medium | 250 < length < 500 |
| H | Heavy | 500 < length |

### 4.2.1. Data Aggregation

The processed sensor data is then aggregated at 1 minute interval to evaluate certain parameters that are average speed of vehicle, total count of vehicles, and lane wise total count of each class of vehicle.

### 4.2.2. Congestion Model

To evaluate the congestion level on the road, it is necessary to make use of an appropriate congestion model which quantifies the congestion status of the road. Among the many types of congestion models explained in section 2.3, this research makes use of empirical model based congestion estimation due to availability of many parameters from the sensor data. The current available congestion models suitable for Indian road conditions differ by the type of input parameter(s) considered for the estimation of congestion. The basic parameters utilized by various different researches are namely: average speed, travel time index, traffic flow. Although most of these studies have promising results, but these major drawback is that they consider only a single parameter for congestion evaluation. As discussed in section 3.2.2, the sensor considered in this research, is capable of providing multiple parameters. Hence it would be unfair to evaluate congestion on the basis of just single parameter. So the focus of this research is on empirical models where more than one input parameters are considered for the purpose of congestion evaluation. Putting light on this problem, Maitra, Sikdar, & Dhingra (1999) proposed a congestion model which considers the relation between operating speed of flow, density of vehicles, capacity of road and the traffic volume. They considered the values for capacity and volume in Passenger Car Unit per Hour (PCUPH). PCUPH is used to estimate the traffic flow rate. It measures the effect of mode of transport on traffic variables like speed, headway and density. For this research, same congestion model will be utilized. The empirical formulas used for the model are explained in (Equation 2), (Equation 3), (Equation 4), and (Equation 5).

(Equation 2) gives a model based relation between the operating speed and the operating volume of the traffic. Since for this research, the volume parameter is unknown but the speed parameter is known, the given equation is reformulated in form of (Equation 3). (Equation 4) evaluates the traffic volume (in PCUPH) at 100% congestion.

$$S = S_f\left[1 - a\left(\frac{V}{C}\right)^{\Sigma p_i m_i}\right] \qquad \text{(Equation 2)}$$

$$V = C\left[\frac{1}{a}\left(1 - \frac{S}{S_f}\right)\right]^{1/\Sigma p_i m_i} \qquad \text{(Equation 3)}$$

$$V_L = \left[\frac{1}{a}\left(1 - \frac{S_L}{S_f}\right)\right]^{1/\Sigma p_i m_i} \qquad \text{(Equation 4)}$$

$$CG_V = \left(\frac{V}{V_L}\right)^{1 + \Sigma p_i m_i} \times 100 \qquad \text{(Equation 5)}$$

Where,
$p_i$ is the proportion of vehicle type $i$ in the stream,
$m_i$ is the contribution of proportion of vehicle type $i$ in the stream,
$S_f$ is the free flow speed of the stream (in km/h),
$S_L$ is the operational speed of traffic at 100% congestion,
$V$ is the operating total traffic volume in PCUPH (Passenger car units per hour),
$V_L$ is the operating total traffic volume at 100% congestion,
$C$ is the capacity of the road section (in PCUPH),
$S$ is the operational speed of the stream (in km/h),
$a$ is the parameter calibrated for the road condition.

The inductive loop sensor used for this research provides values for $p_i$ and $S$. For remaining values, standard values according to Indian / Dehradun roads have been considered which are discussed in Table 4.6.

Table 4.6: Parameters for congestion model

| Parameter | Value | Source |
|---|---|---|
| $\boldsymbol{m_i}$ | $m_i \begin{cases} 0.643, & for\ vehicle\ class\ L \\ 1.257, & for\ vehicle\ class\ M \\ 0.657, & for\ vehicle\ class\ H \end{cases}$ | (Maitra et al., 1999) |
| $\boldsymbol{S_f}$ | 40 Km/hr, Speed limit on road in Dehradun | (Uttarakhand Police, 1988) |
| $\boldsymbol{S_L}$ | 20 Km/hr, Speed at 100% congestion for 4 lane arterial roads | (Maitra et al., 1999) |
| $\boldsymbol{C}$ | 3500 PCUPH, for 4 lane arterial roads | (Congress, 1990) |
| $\boldsymbol{a}$ | 0.6603, for 4 lane arterial roads | (Maitra et al., 1999) |

Maitra et al. (1999) further used these estimated congestion values and divided them into 10 level of service or 10 congestion levels, decided based on the variation of congestion level of their study area. Moreover, Patel & Gundaliya (2016) used the same congestion model but aggregated the congestion percentage values to five different levels. Hence, according to the traffic scenario of this case study, in this research we have aggregated the congestion percentage values to three different congestion levels as specified in Table 4.7.

Table 4.7: Congestion level ranges

| Congestion level | Congestion percentage ($CG_v$) |
|---|---|
| **Low** | $CG_v < 40\%$ |
| **Medium** | $40\% \leq CG_v \leq 80\%$ |
| **High** | $80\% < CG_v$ |

### 4.2.3. Sensor Data storage

The results of congestion level along with the congestion percentage, average speed, count of vehicles and other required attributes are structured in a dictionary format and then stored in a collection in MongoDB.

## 4.3. Data Integration

The results obtained from section 4.1.5 and section 4.2.3 act as input for the data integration model. This section explains how the processed data obtained from the two sources are integrated together. For integration, the research focuses on using MFRI model. K. Runyoro & Ko (2013) suggested that MFRI is effective in traffic level detection and decision making. T. Adetiloye & Awasthi (2017) provided set of rules defined for integration of sensor data with social media data for a short span of time. This research modified the set of rules according to the collected data and provided net set of rules which are explained further in detail in Section 4.3.1. Only those tweets are considered which were in the vicinity of sensors.

### 4.3.1. Rule based integration model

In rule based integration models, human operator's knowledge for integration task is gathered. In conventional rule based model each decision rule is given equal weightage while in fuzzy rule based model a weighing or certainty parameter is used which denotes the strength of each decision rule. Simultaneous triggering of multiple rules are allowed and decision can be made in favour of the rule with greatest strength (Tso & Mather, 2009). On the basis of the available data with providing higher weightage to the sensor results, the provided new set of rules has been shown in Table 4.9. The outputs based on the rules are shown in Table 4.8.

Table 4.8: Rule based data integration of Twitter and Sensor data

| | | **Congestion Level** | | |
|---|---|---|---|---|
| | | **Low** | **Medium** | **High** |
| **Twitter Sentiment** | **Positive** | Possible Low | Medium | High |
| | **Neutral** | Low | Possible Medium | High |
| | **Negative** | Low | Medium | Possible High |

Table 4.9: Proposed rules for integration of road traffic data from the sensors with social media data

**Rule 1:**

If (tweets sentiment is *“**positive**”* and sensor congestion is *“**high**”*)

Then (integrated congestion is *“**high**”)*

**Rule 2:**

If (tweets sentiment is *“**positive**”* and sensor congestion is *“**medium**”*)

Then (integrated congestion is *“**medium**”*)

**Rule 3:**

If (tweets sentiment is *“**positive**”* and sensor congestion is *“**low**”*)

Then (integrated congestion is *“**possible low**”*)

**Rule 4:**

If (tweets sentiment is *“**neutral**”* and sensor congestion is *“**high**”*)

Then (integrated congestion is *“**high**”*)

**Rule 5:**

If (tweets sentiment is *“**neutral**”* and sensor congestion is *“**medium**”*)

Then (integrated congestion is *“**possible medium**”*)

**Rule 6:**

If (tweets sentiment is *“**neutral**”* and sensor congestion is *“**low**”*)

Then (integrated congestion is *“**low**”*)

**Rule 7:**

If (tweets sentiment is *“**negative**”* and sensor congestion is *“**high**”*)

Then (integrated congestion is *“**possible high**”*)

**Rule 8:**

If (tweets sentiment is *“**negative**”* and sensor congestion is *“**medium**”*)

Then (integrated congestion is *“**medium**”*)

**Rule 9:**

If (tweets sentiment is *“**negative**”* and sensor congestion is *“**low**”*)

Then (integrated congestion is *“**low**”*)

## 4.4. Web GIS Dashboard Application

For visualization of results of Twitter based traffic incidents, sensor based traffic congestion and integration model based traffic congestion, a web based GIS application prototype has been developed which is built targeting the transport engineers of Dehradun for better understanding the congestion status. This application provides interactive map based user interface which allows its user to interact with the features to display different visualizations forms of the data. The technologies used for development of the prototype have been mentioned in the Table 2.1. The prototype has been explained in detail in Section 5.4.

# 5. RESULTS & DISCUSSION

## 5.1. Twitter Data Analysis

### 5.1.1. Extracted Twitter data

On querying the Twitter Streaming API with the selected keywords as mentioned in Section 4.1.1, it was observed that the first traffic-related tweet for Dehradun city was posted on 4th April 2009. Since then, a total of 1885 traffic related tweets have been posted till 16th February 2019. For this research, these 1885 tweets were extracted but were used just for the preparation of appropriate training data for classification purpose and were not used further for the integration module. The annual growth pattern is observed for these tweets for past 10 years as is shown in Figure 5.1. This trend displays how the frequency of availability of traffic related tweets had grown in past two years.

This research project began in October 2018, and the traffic-related tweets posted since that time only will be considered for this research. This resulted in a total of 398 tweets from the four and a half month span as shown in Table 5.1. During this span, for the initial four months it was observed that the amount of traffic-related tweets is very less, sometimes even less than a single tweet a day.

It was assessed that the amount of received tweets was not enough to evaluate the congestion. Hence a social experiment was conducted by the author in which alongwith a peer group which consisted of 8 members. In this social experiment, the peers posted tweets from different locations, majorly near the sensors providing an insight of traffic situation on the road at that time at different time of the day. The effect of this social experiment is observed in Table 5.1 with the spike in number of tweets for the month of February, 2019.

Table 5.1: Monthly statistics on total tweets from Dehradun city in past 5 months

| 2018-2019 | Traffic tweets from Dehradun |
|---|---|
| October | 67 |
| November | 31 |
| December | 26 |
| January | 46 |
| February (till 16th Feb) | 228 |
| Total | 398 |

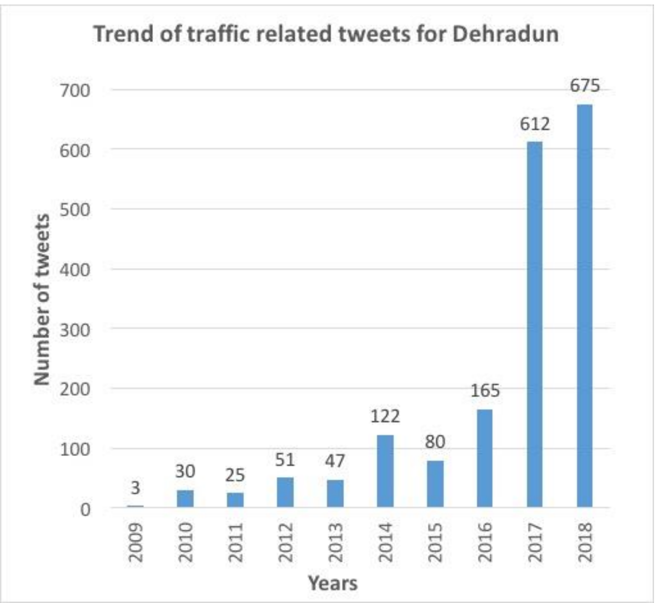


Figure 5.1: Annual Trend of Traffic related tweets in Dehradun city for past 10 years

The hourly trend of tweets distribution for a month and both for a year were observed. The result of the tweets posted at different interval observed for the month of January are shown in Table 5.2. The distribution of the tweets over the different timespans like Morning, Noon, Evening and Night are then observed as shown in Figure 5.2. It can be observed that most of the traffic related tweets are executed during noon time.  Noon time is the school dispersal time as well as office lunch break hour due to which more vehicles are plying on the road. This causes congestion on roads, with increased congestion and hence more people having tendency to tweet about it. This can help come to an inference that Twitter can be used as an efficient data source because the traffic tweet distribution matches the general trend of road traffic congestion.

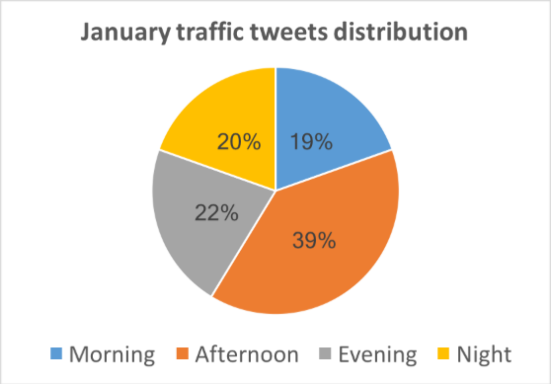


Figure 5.2: Distribution of tweets according to time for January

Table 5.2: Hourly traffic tweets for January 2019

| | Time | Hourly Tweets |
|---|---|---|
| Morning | 5:00 - 6:00 | 1 |
| | 6:00 - 7:00 | 2 |
| | 7:00 - 8:00 | 1 |
| | 8:00 - 9:00 | 0 |
| | 9:00 - 10:00 | 1 |
| | 10:00 - 11:00 | 2 |
| | 11:00 - 12:00 | 2 |
| Noon | 12:00 - 13:00 | 4 |
| | 13:00 - 14:00 | 2 |
| | 14:00 - 15:00 | 4 |
| | 15:00 - 16:00 | 4 |
| | 16:00 - 17:00 | 4 |
| Evening | 17:00 - 18:00 | 2 |
| | 18:00 - 19:00 | 4 |
| | 19:00 - 20:00 | 2 |
| | 20:00 - 21:00 | 1 |
| | 21:00 - 22:00 | 1 |
| Night | 22:00 - 23:00 | 2 |
| | 23:00 - 00:00 | 2 |
| | 0:00 - 1:00 | 5 |
| | 1:00 - 2:00 | 0 |
| | 2:00 - 3:00 | 0 |
| | 3:00 - 4:00 | 0 |
| | 4:00 - 5:00 | 0 |

### 5.1.2. Geocoding Tweets

Out of the 398 tweets which were extracted for this research project, only a few of them tweets were found to be geotagged. These geotagged ones were also the ones which were posted at the time of social experiment. Hence making it a difficult process to directly find out the focal point of the tweet. Hence, to deal with such situation, geocoding services had to be used. For this research, Google Maps based geocoder API was used because of the reliability of its result as well as its seamless support for Hindi language. The API was used to geocode the non-geotagged tweets and reverse geocode the geotagged tweets. This resulted in all the tweets have location as well as coordinates assigned to them as attributes. During this process, the tweets which provided no location information were discarded. Also the tweets where the location was marked as "*Dehradun, Uttarakhand, India*" were removed.

After this process, 159 out of the total 398 tweets were left which proper coordinates and location had attributed. A sample of results obtained after geocoding process are highlighted in Table 5.3.

Table 5.3: Results of Geocoding of tweets

| Language | Type of Tweet | Tweet | Coordinates | Location |
|---|---|---|---|---|
| English | Geotagged Tweet | Pleasant drive at Darshanlal chowk a rare sight… #Dehradun 😋 #Traffic https://t.co/DykIwvdMiA | 30.3229011, 78.0427114 | Pant Rd, Darshan Lal Chowk, Race Course, Dehradun, Uttarakhand 248001, India |
| English | Non geotagged Tweet | Stuck at the Dehradun airport for over an hour.. traffic has been blocked across the highways indefinitely till his highness Rajnath Singh ( BJP) decides to show up at the airport on his way out of the city https://www.facebook.com/1381054401/posts/10212386467041696/ | 30.1949249, 78.1920495 | Dehradun Airport ddn, Airport Road, Dehradun, Uttarakhand 248140, India |
| Hindi | Geotagged Tweet | देहरादून कि पतली गलियां राजधानी के यातायात का भार नहीं झेल पाती। #जाम | 30.3104602, 78.0482533 | Hotel M J Residency, 54, Haridwar Road, Near C.M.I. Hospital, Dehradun Uttarakhand, 248001, India |
| Hindi | Non Geotagged Tweet | # TrafficAlert दिनांक 21.06.2018 को # FRI देहरादून में प्रस्तावित #अंतर्राष्ट्रीय_योग_दिवस कार्यक्रम के दृष्टिगत् यातायात प्लान तैयार किया गया है। कृपया यातायात व्यवस्थापन में # DoonPolice को सहयोग प्रदान करें!pic.twitter.com/lRldKrfQOC | 30.3437685, 77.9995589 | Forest Research Institute, Indian Military Academy, Dehradun, Uttarakhand, India |
| Hinglish | Geotagged Tweet | #Modi aaj sheher me! Pure Dehradun ka traffic idhar udhar bhaag rha hai! (~_~) Parade ground k 2km tk koi gaadi allowed nahi... Saare raste blocked # Modi | 30.324502, 78.0484316 | Race Course, Dehradun, Uttarakhand 248001, India |
| Hinglish | Non Geotagged Tweet | @ DehradunSsp . @ DehradunDm . @ HighUttarakhand . @ SupremeCourtIND. @ tsrawatbjp . @ narendramodi . Brahman Mohalla, Raipur, Dehra Dun me dadagiri kar atikraman ko nahi todne diya jaa raha hai. Ye sadak Maldevta jaane ke liye main maarg h aur hamesha traffic jam rehta h! https://twitter.com/JaiBharat07/status/1021438539890929665 | 30.3363803, 78.1442051 | Maldevta Road, Uttarakhand, India |

### 5.1.3. Sentiment analysis

The resultant tweets after geotagging go through the process of sentiment analysis in which the sentiment associated with the tweet is found out on the basis of Afinn score. A sample of the result obtained after sentiment analysis is shown in Table 5.4. From the result, it can be observed some of the tweets in Hindi are showing a neutral Afinn score. This was due to the fact that the sentiment analysis method use had limited scope for Hindi language. New words related to traffic situations were introduced in the Afinn library of 3382 words. To cover the whole scope of language, dedicated Hindi dictionary based terms can be used implied in future. The results obtained from this analysis are further used to mark the congestion status on the road. Although its weightage is kept low in comparison to the sensor result.

Table 5.4: Result of Sentiment analysis

| Tweet | Afinn Score | Sentiment |
|---|---|---|
| Pleasant drive at Darshanlal chowk a rare sight… #Dehradun #Traffic 😋 https://t.co/DykIwvdMiA | 2.0 | Positive |
| Stuck at the Dehradun airport for over an hour.. traffic has been blocked across the highways indefinitely till his highness Rajnath Singh ( BJP) decides to show up at the airport on his way out of the city...... https://www.facebook.com/1381054401/posts/10212386467041696/ … | -4.0 | Negative |
| देहरादून कि पतली गलियां राजधानी के यातायात का भार नहीं झेल पाती। #जाम | -2.0 | Negative |
| # TrafficAlert दिनांक 21.06.2018 को # FRI देहरादून में प्रस्तावित #अंतर्राष्ट्रीय_योग_दिवस कार्यक्रम के दृष्टिगत् यातायात प्लान तैयार किया गया है। कृपया यातायात व्यवस्थापन में # DoonPolice को सहयोग प्रदान करें!pic.twitter.com/lRldKrfQOC | 0.0 | Neutral |
| #Modi aaj sheher me! Pure Dehradun ka traffic idhar udhar bhaag rha hai! (~_~) Parade ground k 2km tk koi gaadi allowed nahi... Saare raste blocked # Modi | -2.0 | Negative |
| @ DehradunSsp . @ DehradunDm . @ HighUttarakhand . @ SupremeCourtIND . @ tsrawatbjp . @ narendramodi . @ CMuttarakhand Brahman Mohalla, Raipur, Dehra Dun me dadagiri kar atikraman ko nahi todne diya jaa raha hai. Ye sadak Maldevta jaane ke liye main maarg h aur hamesha traffic jam rehta h!!! | -2.0 | Negative |

### 5.1.4. Traffic incident classification

The geocoded tweets with sentiment attached to them are now processed for classifying them the type of incident accordingly. The process of classification as discussed in Section 4.1.4 starts with the processing the textual tweets in order to convert them in numeric form. The 1885 traffic related tweets which were initially collected from the data extraction process are now used for this process for generation of training dataset. These tweets are manually labelled into 7 different classes as mentioned in Table 4.4. Two different samples highlighting how the text is processed before the classification for English and Hindi tweets is shown in Table 5.5 and Table 5.6. From the output it can be observed that wherever possible, the words are converted to their shortest form (lemmatization) so that if different forms of same words are not marked differently.

Table 5.5: Text Processing Result of English tweet

| | |
|---|---|
| **Original Text** | Slow traffic at Dilaram Chowk due to a collision of 3 cars 😓 #accident pic.twitter.com/4srfhmemo3 |
| **Lower Case** | slow traffic at dilaram chowk due to a collision of 3 cars 😓 #accident pic.twitter.com/4srfhmemo3 |
| **URL's removed** | slow traffic at dilaram chowk due to a collision of 3 cars 😓#accident |
| **Numbers, Emojis & Symbols removed** | slow traffic at dilaram chowk due to a collision of  cars  accident |
| **Text divided into Tokens** | ['slow', 'traffic', 'at', 'dilaram', 'chowk', 'due', 'to', 'a', 'collision', 'of', 'cars', 'accident'] |
| **Remove Stop words and Lemmatization** | ['slow', 'dilaram', 'due', 'collide', 'car', 'accident'] |

Table 5.6: Text Processing Result of Hindi tweet

| | |
|---|---|
| **Original Text** | # स्वतन्त्रता दिवस के अवसर पर # यातायात एवं डायवर्ट व्यवस्था http://www.newspost.live/dehradun-traffic-divert-route-plan-for-tomorrow/ … via @ News Post @ dehradunSsp # independenceday2017 |
| **Lower Case** | # स्वतन्त्रता दिवस के अवसर पर # यातायात एवं डायवर्ट व्यवस्था http://www.newspost.live/dehradun-traffic-divert-route-plan-for-tomorrow/ … via @ news post @ dehradunssp # independenceday2017 |
| **URL's removed** | # स्वतन्त्रता दिवस के अवसर पर # यातायात एवं डायवर्ट व्यवस्था … via @ news post @ dehradunssp # independenceday2017 |
| **Numbers, Emojis & Symbols removed** | स्वतन्त्रता दिवस के अवसर पर  यातायात एवं डायवर्‌ट व्यवस्था  via  news post dehradunssp  independenceday |
| **Text divided into Tokens** | ['स्वतन्त्रता', 'दिवस', 'के', 'अवसर', 'पर', 'यातायात', 'एवं', 'डायवर्ट', 'व्यवस्था', 'via', 'news', 'post', 'dehradunssp', 'independenceday'] |
| **Remove Stop words and Lemmatization** | ['स्वतन्त्रता', 'दिवस', 'अवसर', 'यातायात', 'डायवर्ट', 'व्यवस्था', 'news', 'post', 'independenceday'] |

For the classification of the tweets NBM classification algorithm has been used. The dataset of 1885 tweets after processing is split into training set and testing set using the standard practice ratio of 70% and 30%, hence randomly assigning 1319 tweets for training and 566 tweets for testing. The result obtained after classification is highlighted in Table 5.7 in form of confusion matrix. From the result of the matrix, the accuracy of the classifier is evaluated which is obtained as **79.15 %**. It is observed that the maximum accuracy was obtained for class 'weather' in which 73 out of 83 testing samples were correctly classified. While remaining of them were misclassified with other classes. It is noted that 'weather' and 'congestion' have mostly misclassified with each other. This is because of the fact that the state of 'congestion' and the state of 'weather' are described with similar words. It can also be observed that class 'congestion' and 'other' are most number of samples as well as misclassifications. In case of 'congestion' it is because many common words have gone into training of 'congestion' as well as remaining classes, which makes the misclassification more frequent. This can be improved by providing more number of stop words according to the purpose of classification. Class 'other' was trained with all sorts of tweets which were non relevant to any other class which creates a huge range of misclassification for 'other' class. The accuracy obtained with NBM is found to be consistent with the other studies where the same algorithm has been used for textual classification (Kiilu et al., 2018; K. Lee et al., 2011). Hence for the identification of type of event this training set is accepted and used for all future predictions of the incidents.

Table 5.7: Confusion Matrix NBM Classifier

| | | Classified Data | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **congestion** | **accident** | **diversion** | **construction** | **weather** | **event** | **other** | **Total** | **Producer Accuracy (%)** |
| **Testing data** | **congestion** | 126 | 5 | 2 | 2 | 5 | 6 | 6 | 152 | **82.89** |
| | **accident** | 2 | 51 | 0 | 3 | 2 | 1 | 1 | 60 | **85.00** |
| | **diversion** | 1 | 0 | 33 | 2 | 1 | 2 | 4 | 43 | **76.74** |
| | **construction** | 3 | 2 | 3 | 35 | 4 | 0 | 4 | 51 | **68.63** |
| | **weather** | 5 | 1 | 0 | 2 | 73 | 0 | 2 | 83 | **87.95** |
| | **event** | 7 | 2 | 3 | 1 | 0 | 41 | 3 | 57 | **71.93** |
| | **other** | 10 | 4 | 4 | 2 | 7 | 4 | 89 | 120 | **74.17** |
| | **Total** | 154 | 65 | 45 | 47 | 92 | 54 | 109 | 566 | |
| | **User Accuracy (%)** | **81.82** | **78.46** | **73.33** | **74.47** | **79.35** | **75.93** | **81.65** | | |

| Overall Accuracy | Kappa Score |
|---|---|
| 79.15 % | 0.66 |

A sample of results obtained after the classification of the geocoded tweets is shown in Table 5.8. It highlights samples from all the classes.

Table 5.8: Results showing classified traffic incidents

| **Tweet** | **Incident** |
|---|---|
| Bad roads, traffic jams and overflowing drains have made life miserable for the residents of # TransportNagar # Dehradun | congestion |
| @ DehradunSsp .Madem,Huge traffic jam at kargi Chowk near SBI bank pic.twitter.com/vMstpEXPY2 | congestion |
| Traffic chaos continues despite free parking lots outside schools in Dehradun - Hindustan Times http://ift.tt/2BaMALU # Dehradun # news | congestion |
| # Dehradun डाटकाली मंदिर के पास टू-लेन टनल का निर्माण कार्य शुरू, टनल बनने के बाद जाम से लोगों को मिलेगी निजात @ DehradunDm @ madankaushikbjppic.twitter.com/vj3iHTe15f | construction |
| Under-construction flyover adds to traffic woes: DEHRADUN: The residents of Clement Town on the... http://binged.it/1RKZXEd # driving # news | construction |
| # NHAI Dear Sir without constructing Service lane Traffic diverted to muddy road at NH72 Mohkam pur Rly Crossing at Dehradun, nobody bothers pic.twitter.com/fLKkhDOCkn | construction |
| Tumko Jo karna hai Karo...but please .. jab Dehradun aate ho to airport ka traffic block mat karo.. Dehradun airport is adjacent to jallygrant hospital..tumhari visit main ...raat 9:30 baje 40min ka road block thaa...aaj vaikaih Naidu ki waja se 30min block thaa.. | diversion |
| Traffic Diversion for Passing Out Parade, IMA Dehradun, Pioneer @ TheDailyPioneer http://www.dailypioneer.com/state-editions/dehradun/police-all-set-for-a-smooth-traffic-flow-during-pop.html … | diversion |
| Modi in town today! Entire Dehradun traffic has been stranded! (~_~) No entry zone within 2km. radius of Parade Ground... # Modi | event |
| देहरादून-शहर की यातायात व्यवस्था लड़खड़ाई रिस्पना पुल पर जाम में फंसे एसएसपी दून जगह-जगह होली के त्योहार के चलते जाम | event |
| Rains, landslides obstruct traffic on highways in U'khand: Dehradun, Aug 11: Light to medium rain in various... http://bit.ly/1oEGsPn | weather |
| # Dehradun भारी बारिश से जन जीवन अस्त-व्यस्त,दून में कई जगह तालाब में तब्दील हुई सड़कें,रिस्पना,राजपुर रोड पर यातायात हुआ ठप pic.twitter.com/pmSpcXBFaq | weather |
| Landslides on Doon-Mussoorie road disrupt traffic, commuters hassled - Times of India https://ift.tt/2mbrpRi # Dehradun # news | weather |
| देहरादून मसूरी रोड पर भूस्खलन ,यातायात हुआ बाधित https://youtu.be/OFXEwQ7X9ZM via @ YouTube | weather |
| Lokeshwer Singh transferred as SP Traffic Dehradun District \| Uttarakhand Police http://www.bureaucracynews.com/lokeshwer-singh-transferred-as-sp-traffic-dehradun-district-uttrakhand-police/ … | other |
| It's much easier to double your business by doubling your conversion rate than by doubling your traffic." http://www.orbosyscrop.com \| info@orbosyscorp.com \| +91-9520999099 # DigitalMarketing # digitalmarketingtraining # Entrepreneurs # startup # job # webdesign # web # dehradun # businesspic.twitter.com/LaOL4foldP | other |

### 5.1.5. Twitter Data Storage

The final result obtained after the filtering, geocoding, sentiment analysis and classification of the tweets are stored in dictionary format in MongoDB as shown in Table 5.9.

Table 5.9: Twitter data storage

```
{
   "_id" : 5c777ddd5fc21d0dbc444bdb",
   "type" : "Feature",
   "properties" : {
      "type" : "Non Geotagged Tweet",
      "date" : "10/06/10",
      "time" : "17:57",
      "user" : "wanderabyss",
      "tweet" : "driving back to dehradun.. traffic is like hell in doiwala.. tiny little road & every one just wan to chip in...",
      "incident" : "congestion",
      "sentiment" : "Negative",
      "location" : "Doiwala, Uttarakhand 248140, India",
      "latlng" : "30.1734361, 78.1250848",
      "id" : "15846751503",
      "datetime" : “2010-06-10T17:57:00.000+0000"
   },
   "geometry" : {
      "type" : "Point",
      "coordinates" : [
         78.1250848,
         30.1734361
      ]
   }
}
```

## 5.2. Sensor Data Analysis

The raw data available in textual format as well as XML format as discussed in Section 3.2.2 is processed by method explained in Section 4.2.1 so as to first aggregate the data in per minute interval then evaluate the congestion percentage and congestion level using the model explained in Section 4.2.2 and finally store the output in MongoDB as shown in Section 4.2.3. A sample output of the saved result in tabular form is highlighted in Table 5.10.

Table 5.10: Result obtained applying the congestion model on aggregated sensor data

| Hour | Min | Direction | Density | Average Speed | Count C1 | Count C2 | Count C3 | Speed C1 | Speed C2 | Speed C3 | Congestion | Congestion Level |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 11 | 15 | 2 | 19 | 27.42 | 12 | 6 | 1 | 23.91 | 28.16 | 65 | 56.29 | medium |
| 11 | 16 | 2 | 36 | 37.5 | 19 | 12 | 5 | 22.78 | 29.83 | 111.8 | 20.99 | low |
| 11 | 17 | 2 | 25 | 20.08 | 10 | 15 | 0 | 20.90 | 19.53 | 0 | 94.06 | high |
| 11 | 18 | 2 | 37 | 26.72 | 25 | 9 | 3 | 24.44 | 26.44 | 46.66 | 58.67 | medium |
| 11 | 19 | 2 | 31 | 24.25 | 13 | 15 | 3 | 16.76 | 21.86 | 68.66 | 72.45 | medium |
| 11 | 20 | 2 | 34 | 28.61 | 24 | 9 | 1 | 24.75 | 28.22 | 125 | 50.44 | medium |

### 5.2.1. Sensor data Storage

After the steps of processing, aggregation and congestion modelling, the sensor data is stored within MongoDB in dictionary formats as shown in Table 5.11.

Table 5.11: Sensor data storage

```
{
        "date": 10/12/18,
        "time": 17:57,
        “sensor”: Astley Hall,
        "direction": 2,
        "lane": 3,
        "total_count": 24,
        "averagespeed": 18,
        "count_c1": 9,
        "count_c2": 10,
        "count_c3": 5,
        "congestion": 94.34%,
        "congestionlevel": High
}
```

## 5.3. Data Integration

### 5.3.1. Tweets relevant for integration

So as to filter the relevant tweets which were used to integrate with the sensors a bounding box was defined over each sensor with its extent covering the road segment which is covered by the sensor. All the tweets which were lying within this bounding box are selected for the integrating with the sensor data. Figure 5.3 shows the bounding box around the sensor locations. The filtered tweets are then further sent to the integration model.

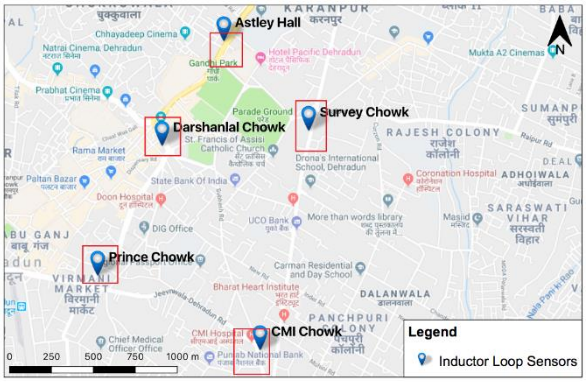


Figure 5.3: Bounding box highlighting area considered per sensor to capture tweets

### 5.3.2. Data storage

The resultant of integration of data are added to the same collection as the sensor in MongoDB and the new parameters are appended in the dictionary as shown in Table 5.12.

Table 5.12: Integrated data storage

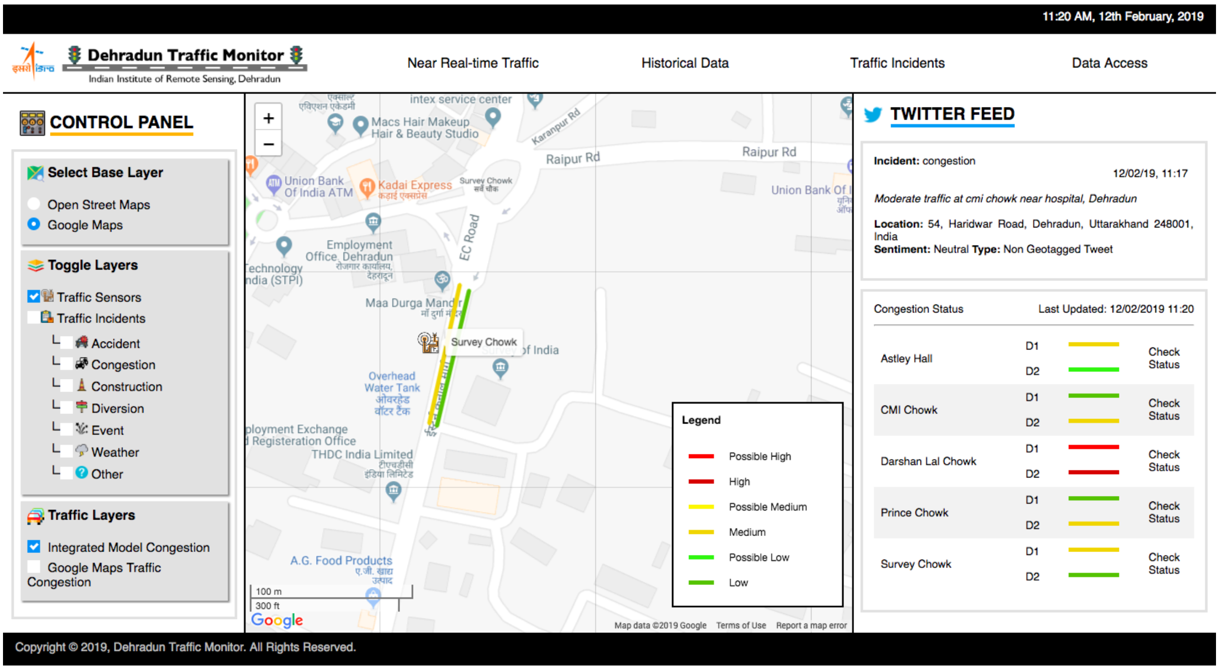
```
Integrated
{
   "_id" : ObjectId("5c78bdccae75e860bbbe2206"),
   "date" : "12/02/19",
   "time" : "12:49",
   "averagespeed" : 24.93,
   "congestion" : 68.75,
   "congestionlevel" : "medium",
   "count_c1" : NumberInt(14),
   "count_c2" : NumberInt(11),
   "count_c3" : NumberInt(2),
   "datetime" : ISODate("2019-02-12T12:49:00.000+0000"),
   "density" : NumberInt(27),
   "direction" : NumberInt(2),
   "hour" : NumberInt(12),
   "integrated_congestion" : "possible_medium",
   "lane" : NumberInt(3),
   "min" : NumberInt(49),
   "sensor" : "CMI CHOWK",
   "speed_c1" : 25.36,
   "speed_c2" : 24.91,
   "speed_c3" : 22.0,
   "incident" : "12:49",
   "sentiment" : "Neutral",
   "tweet" : "Unmanaged traffic cause regular traffic issue at CMI chowk dehradun",
   "user" : "utsav1995s"
}
```

## 5.4. Web GIS Dashboard Application

### 5.4.1. Basic framework

As discussed in Section 4.4, a basic framework of the prototype application was built in form of a Web GIS dashboard application. This dashboard consisted of mainly four modules serving different purpose each. All the modules have a common header, footer and a menu. Rest of the content differ in each page. The resultant dashboard along with all the content pages are discussed in further sections.

#### 5.4.1.1. Near Real time traffic map


11:20 AM, 12th February, 2019
Dehradun Traffic Monitor
Indian Institute of Remote Sensing, Dehradun
Near Real-time Traffic
Historical Data
Traffic Incidents
Data Access
+
−
Macs Hair Makeup Hair & Beauty Studio
Union Bank Of India ATM
Kadai Express
Survey Chowk
Employment Office, Dehradun
EC Road
Raipur Rd
Karanpur Rd
Maa Durga Mandir
Overhead Water Tank
THDC India Limited
A.G. Food Products
100 m
300 ft
Google
Map data ©2019 Google Terms of Use Report a map error
Copyright © 2019, Dehradun Traffic Monitor. All Rights Reserved.

CONTROL PANEL

Select Base Layer

○ Open Street Maps
● Google Maps

Toggle Layers

☑ Traffic Sensors
□ Traffic Incidents
- □ Accident
- □ Congestion
- □ Construction
- □ Diversion
- □ Event
- □ Weather
- □ Other

Traffic Layers

☑ Integrated Model Congestion
□ Google Maps Traffic Congestion

Legend

- 🟥 Possible High
- 🟥 High
- 🟨 Possible Medium
- 🟨 Medium
- 🟩 Possible Low
- 🟩 Low

TWITTER FEED

**Incident:** congestion
12/02/19, 11:17

*Moderate traffic at cmi chowk near hospital, Dehradun*

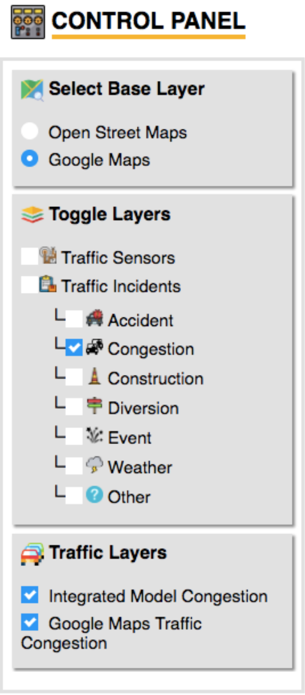
**Location:** 54, Haridwar Road, Dehradun, Uttarakhand 248001, India
**Sentiment:** Neutral **Type:** Non Geotagged Tweet

| Congestion Status | Last Updated: 12/02/2019 11:20 | | |
|---|---|---|---|
| Astley Hall | D1 | 🟨 | Check Status |
| | D2 | 🟩 | |
| CMI Chowk | D1 | 🟩 | Check Status |
| | D2 | 🟨 | |
| Darshan Lal Chowk | D1 | 🟥 | Check Status |
| | D2 | 🟥 | |
| Prince Chowk | D1 | 🟩 | Check Status |
| | D2 | 🟨 | |
| Survey Chowk | D1 | 🟨 | Check Status |
| | D2 | 🟩 | |

Figure 5.4: Dashboard showing near real time traffic situation

This module provides near real time congestion status of the roads segments attached to the sensors as well as traffic incident updates using the results of integration. This module is divided into three panels as shown in Figure 5.4. The left most panel provides controls to the user so as to modify the visual content available on the map panel as shown in Figure 5.5. While the central panel holds the map and displays all the vector layers for incidents (points) & road segments with sensors (poly lines). The panel on the right hand side shows the total incidents reported that day via Twitter as well as the current congestion level at the sensors as shown in Figure 5.6. The color coding used for the congestion levels has been shown in Figure 5.7. The near real time traffic map provides an insight of the incident(s) that were reported for the current date as well as provide per minute update of the traffic situation at every sensor location considering the result of the integration model. The near real time functionality is added by making use of JavaScript based setInterval function which allows to call a function after every fixed interval of time. So the Python script written for Twitter data extraction and processing as well as the Sensor data processing script works in background updating the MongoDB database every minute. While at the same time, PHP makes call to MongoDB to extract latest entry and refreshes the map elements.

Two different base maps are provided, i.e. Open Street Map (OSM) and Google Maps which can be toggled by using the radio button functionality in the Control Panel section. Google Maps has its own near real time updated traffic layer which has can also be toggled so as to compare the congestion values of google maps vs the sensor.



CONTROL PANEL

**Select Base Layer**

○ Open Street Maps
● Google Maps

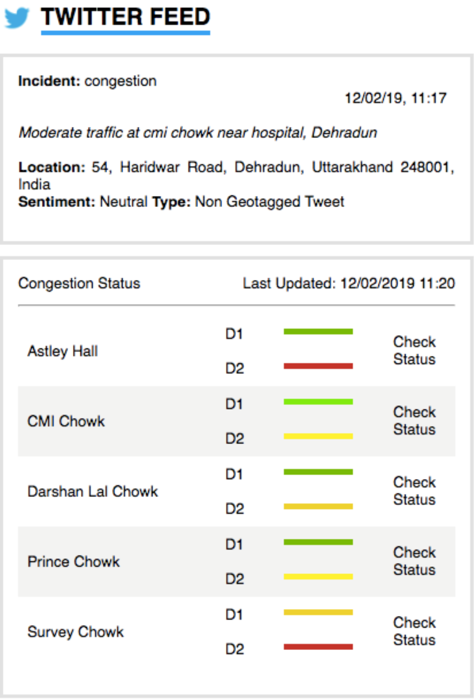
**Toggle Layers**

□ Traffic Sensors
□ Traffic Incidents
└ □ Accident
└ ☑ Congestion
└ □ Construction
└ □ Diversion
└ □ Event
└ □ Weather
└ □ Other

**Traffic Layers**

☑ Integrated Model Congestion
☑ Google Maps Traffic Congestion

Figure 5.5: Near real time dashboard control panel



TWITTER FEED

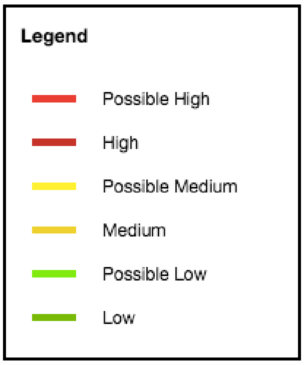
**Incident:** congestion
12/02/19, 11:17

*Moderate traffic at cmi chowk near hospital, Dehradun*

**Location:** 54, Haridwar Road, Dehradun, Uttarakhand 248001, India
**Sentiment:** Neutral **Type:** Non Geotagged Tweet

| Congestion Status | | Last Updated: 12/02/2019 11:20 | |
|---|---|---|---|
| Astley Hall | D1 | 🟩 | Check Status |
| | D2 | 🟥 | |
| CMI Chowk | D1 | 🟩 | Check Status |
| | D2 | 🟨 | |
| Darshan Lal Chowk | D1 | 🟩 | Check Status |
| | D2 | 🟨 | |
| Prince Chowk | D1 | 🟩 | Check Status |
| | D2 | 🟨 | |
| Survey Chowk | D1 | 🟨 | Check Status |
| | D2 | 🟥 | |

Figure 5.6: Near real time result

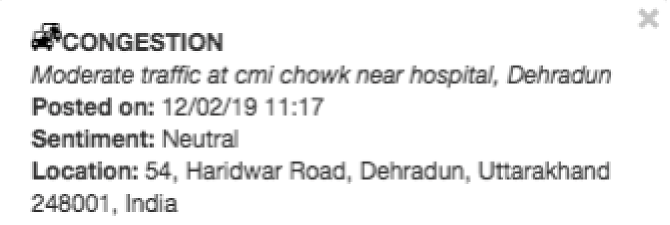


**Legend**

| | |
|---|---|
| 🟥 | Possible High |
| 🟥 | High |
| 🟨 | Possible Medium |
| 🟨 | Medium |
| 🟩 | Possible Low |
| 🟩 | Low |

Figure 5.7: Legend for different congestion levels

The incidents reported by tweets are captured in form of point layers displayed using seven different icons for the seven different classes. All the incident point layers have a functionality of showing the attributes associated with the point layer on clicking the point layer, as shown in figure

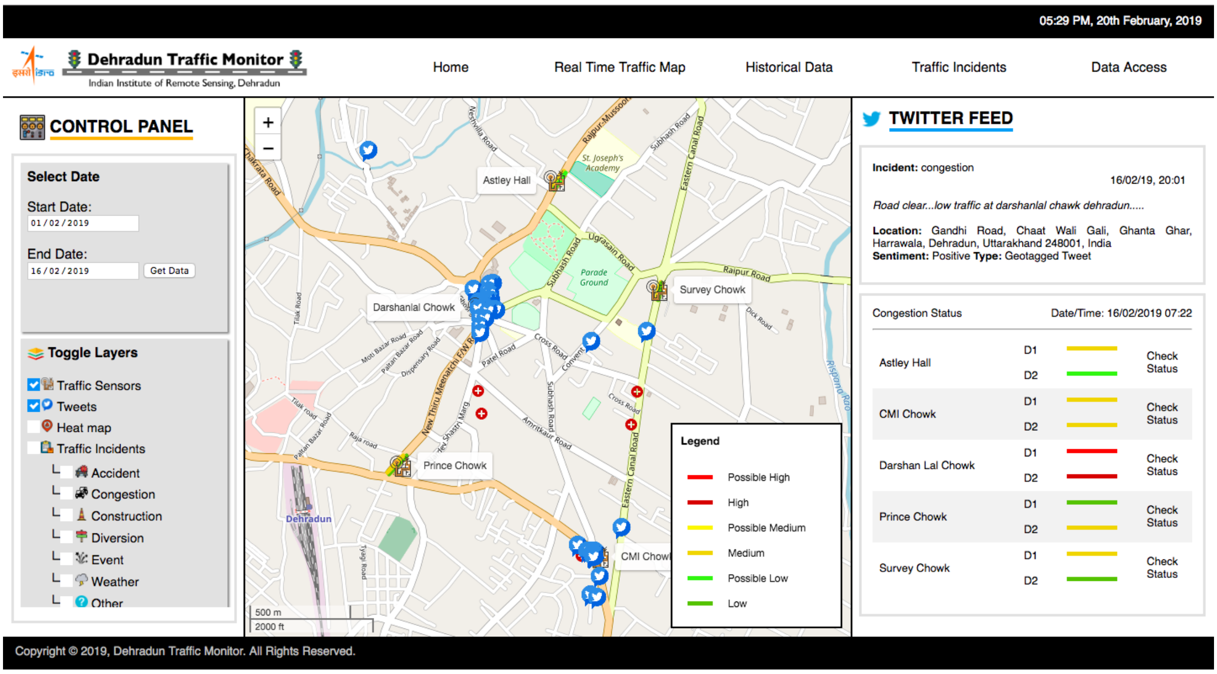


CONGESTION
*Moderate traffic at cmi chowk near hospital, Dehradun*
**Posted on:** 12/02/19 11:17
**Sentiment:** Neutral
**Location:** 54, Haridwar Road, Dehradun, Uttarakhand 248001, India

Figure 5.8: Attributes assigned to each point layer (traffic incident)

### 5.4.1.2. Historical traffic map

This module visually looks similar to real time module but has a few different functionalities. It has some changes in the control panel where the user can select the start and end date. It has a time slider which on scrolling provides an insight of changes in congestion levels as well as by time shows the different incidents which occur with respect to time. This module also allows the user to view all the tweets together and as well as creates a heat map of all the points highlighting the region where the frequency of tweets is higher.Figure 5.9 and Figure 5.10 show the module showing all tweets and heat map.



Select Date

Start Date: 01/02/2019

End Date: 16/02/2019 Get Data

Toggle Layers

- ☑ Traffic Sensors
- ☑ Tweets
- □ Heat map
- □ Traffic Incidents
  - □ Accident
  - □ Congestion
  - □ Construction
  - □ Diversion
  - □ Event
  - □ Weather
  - □ Other

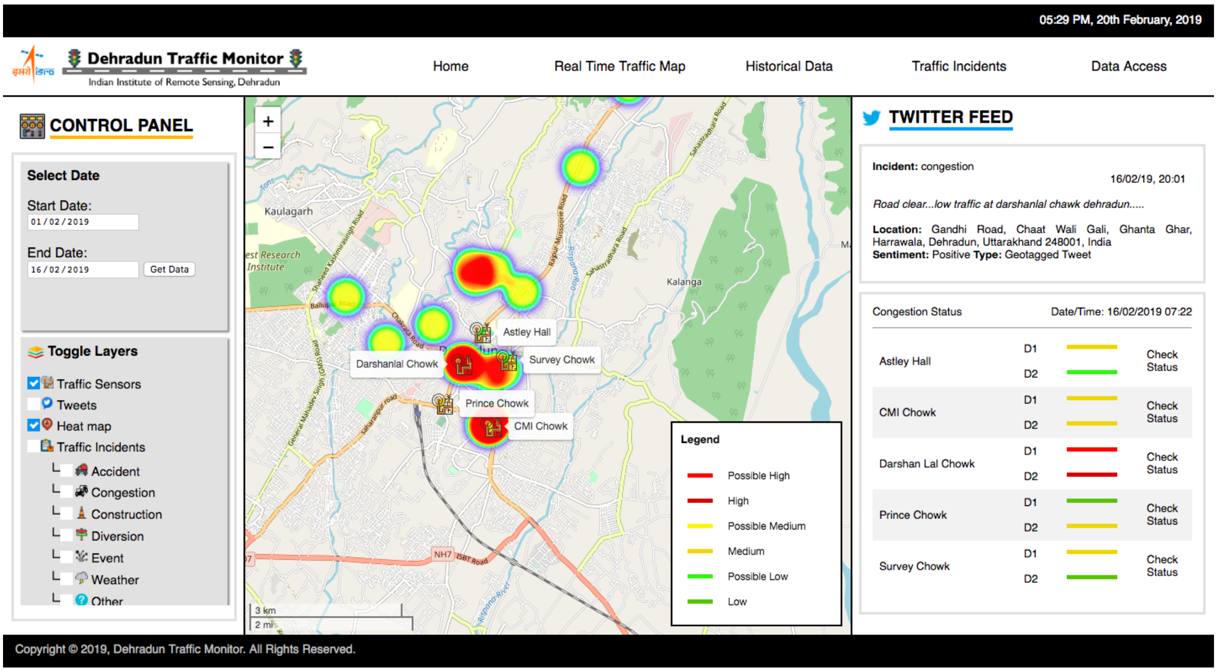
**Incident:** congestion 16/02/19, 20:01

*Road clear...low traffic at darshanlal chawk dehradun.....*

**Location:** Gandhi Road, Chaat Wali Gali, Ghanta Ghar, Harrawala, Dehradun, Uttarakhand 248001, India
**Sentiment:** Positive **Type:** Geotagged Tweet

| Congestion Status | | | Date/Time: 16/02/2019 07:22 |
|---|---|---|---|
| Astley Hall | D1 | 🟨 | Check Status |
| | D2 | 🟩 | |
| CMI Chowk | D1 | 🟨 | Check Status |
| | D2 | 🟨 | |
| Darshan Lal Chowk | D1 | 🟥 | Check Status |
| | D2 | 🟥 | |
| Prince Chowk | D1 | 🟩 | Check Status |
| | D2 | 🟨 | |
| Survey Chowk | D1 | 🟨 | Check Status |
| | D2 | 🟩 | |

Figure 5.10: All tweets shown on Historical map



Select Date

Start Date: 01/02/2019

End Date: 16/02/2019 Get Data

Toggle Layers

- ☑ Traffic Sensors
- Tweets
- ☑ Heat map
- □ Traffic Incidents
  - □ Accident
  - □ Congestion
  - □ Construction
  - □ Diversion
  - □ Event
  - □ Weather
  - □ Other

**Incident:** congestion 16/02/19, 20:01

*Road clear...low traffic at darshanlal chawk dehradun.....*

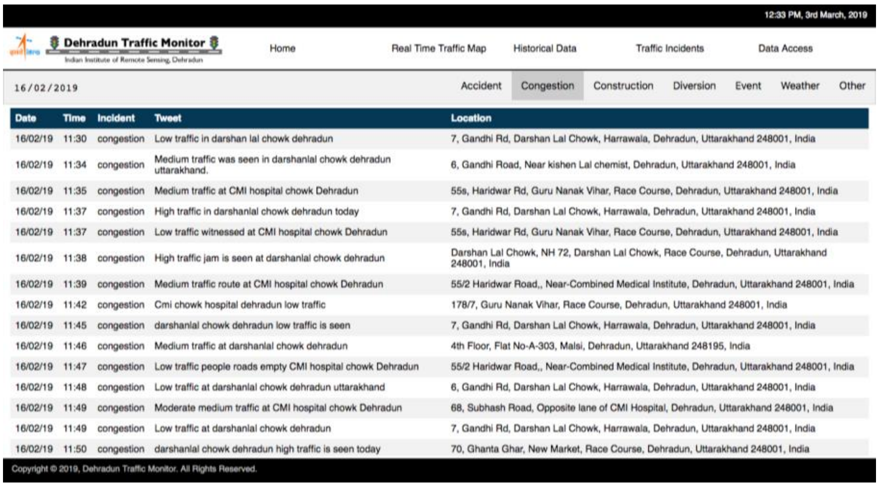
**Location:** Gandhi Road, Chaat Wali Gali, Ghanta Ghar, Harrawala, Dehradun, Uttarakhand 248001, India
**Sentiment:** Positive **Type:** Geotagged Tweet

| Congestion Status | | | Date/Time: 16/02/2019 07:22 |
|---|---|---|---|
| Astley Hall | D1 | 🟨 | Check Status |
| | D2 | 🟩 | |
| CMI Chowk | D1 | 🟨 | Check Status |
| | D2 | 🟨 | |
| Darshan Lal Chowk | D1 | 🟥 | Check Status |
| | D2 | 🟥 | |
| Prince Chowk | D1 | 🟩 | Check Status |
| | D2 | 🟨 | |
| Survey Chowk | D1 | 🟨 | Check Status |
| | D2 | 🟩 | |

Figure 5.9: Historical Map showing heat map

### 5.4.1.3. Traffic incidents

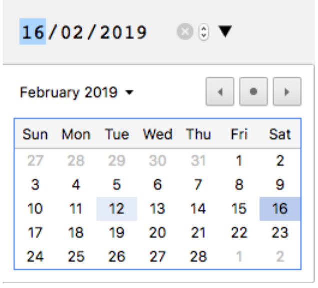
This module comprises provides the user the access to all the traffic incidents that happened on a single day in a tabular fashion. The user can sort through the table on the basis of different attributes that the incident possess. Figure 5.11 and Figure 5.12 provide a glimpse of the Module showing Traffic Incidents.



| Date | Time | Incident | Tweet | Location |
|---|---|---|---|---|
| 16/02/19 | 11:30 | congestion | Low traffic in darshan lal chowk dehradun | 7, Gandhi Rd, Darshan Lal Chowk, Harrawala, Dehradun, Uttarakhand 248001, India |
| 16/02/19 | 11:34 | congestion | Medium traffic was seen in darshanlal chowk dehradun uttarakhand. | 6, Gandhi Road, Near kishen Lal chemist, Dehradun, Uttarakhand 248001, India |
| 16/02/19 | 11:35 | congestion | Medium traffic at CMI hospital chowk Dehradun | 55s, Haridwar Rd, Guru Nanak Vihar, Race Course, Dehradun, Uttarakhand 248001, India |
| 16/02/19 | 11:37 | congestion | High traffic in darshanlal chowk dehradun today | 7, Gandhi Rd, Darshan Lal Chowk, Harrawala, Dehradun, Uttarakhand 248001, India |
| 16/02/19 | 11:37 | congestion | Low traffic witnessed at CMI hospital chowk Dehradun | 55s, Haridwar Rd, Guru Nanak Vihar, Race Course, Dehradun, Uttarakhand 248001, India |
| 16/02/19 | 11:38 | congestion | High traffic jam is seen at darshanlal chowk dehradun | Darshan Lal Chowk, NH 72, Darshan Lal Chowk, Race Course, Dehradun, Uttarakhand 248001, India |
| 16/02/19 | 11:39 | congestion | Medium traffic route at CMI hospital chowk Dehradun | 55/2 Haridwar Road,, Near-Combined Medical Institute, Dehradun, Uttarakhand 248001, India |
| 16/02/19 | 11:42 | congestion | Cmi chowk hospital dehradun low traffic | 178/7, Guru Nanak Vihar, Race Course, Dehradun, Uttarakhand 248001, India |
| 16/02/19 | 11:45 | congestion | darshanlal chowk dehradun low traffic is seen | 7, Gandhi Rd, Darshan Lal Chowk, Harrawala, Dehradun, Uttarakhand 248001, India |
| 16/02/19 | 11:46 | congestion | Medium traffic at darshanlal chowk dehradun | 4th Floor, Flat No-A-303, Malsi, Dehradun, Uttarakhand 248195, India |
| 16/02/19 | 11:47 | congestion | Low traffic people roads empty CMI hospital chowk Dehradun | 55/2 Haridwar Road,, Near-Combined Medical Institute, Dehradun, Uttarakhand 248001, India |
| 16/02/19 | 11:48 | congestion | Low traffic at darshanlal chowk dehradun uttarakhand | 6, Gandhi Rd, Darshan Lal Chowk, Harrawala, Dehradun, Uttarakhand 248001, India |
| 16/02/19 | 11:49 | congestion | Moderate medium traffic at CMI hospital chowk Dehradun | 68, Subhash Road, Opposite lane of CMI Hospital, Dehradun, Uttarakhand 248001, India |
| 16/02/19 | 11:49 | congestion | Low traffic at darshanlal chowk dehradun | 7, Gandhi Rd, Darshan Lal Chowk, Harrawala, Dehradun, Uttarakhand 248001, India |
| 16/02/19 | 11:50 | congestion | darshanlal chowk dehradun high traffic is seen today | 70, Ghanta Ghar, New Market, Race Course, Dehradun, Uttarakhand 248001, India |

Copyright © 2019, Dehradun Traffic Monitor. All Rights Reserved.

Figure 5.12: Prototype Module showing Traffic incidents

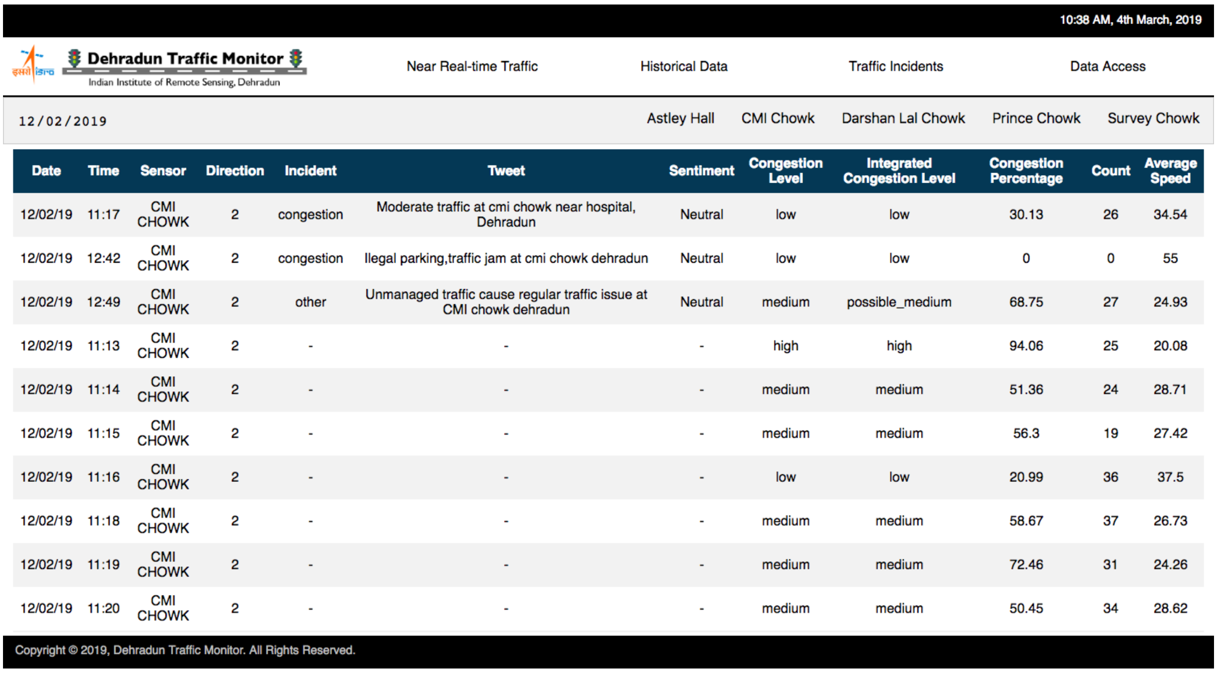


| Sun | Mon | Tue | Wed | Thu | Fri | Sat |
|---|---|---|---|---|---|---|
| 27 | 28 | 29 | 30 | 31 | 1 | 2 |
| 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| 17 | 18 | 19 | 20 | 21 | 22 | 23 |
| 24 | 25 | 26 | 27 | 28 | 1 | 2 |

Figure 5.11: Calendar date drop down

### 5.4.1.4. Data access

This module is similar to traffic Incident module. It provides the user the access to location and date wise historical data of the sensor. This comprises of data in tabular form and highlights the result of integration module. The user can filter the data according to the attribute of the data. For instance it can sort the data according to different attributes like average speed, congestion percent, count of vehicles etc. A glimpse of this Module has been shown in Figure 5.13.



| Date | Time | Sensor | Direction | Incident | Tweet | Sentiment | Congestion Level | Integrated Congestion Level | Congestion Percentage | Count | Average Speed |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 12/02/19 | 11:17 | CMI CHOWK | 2 | congestion | Moderate traffic at cmi chowk near hospital, Dehradun | Neutral | low | low | 30.13 | 26 | 34.54 |
| 12/02/19 | 12:42 | CMI CHOWK | 2 | congestion | llegal parking,traffic jam at cmi chowk dehradun | Neutral | low | low | 0 | 0 | 55 |
| 12/02/19 | 12:49 | CMI CHOWK | 2 | other | Unmanaged traffic cause regular traffic issue at CMI chowk dehradun | Neutral | medium | possible_medium | 68.75 | 27 | 24.93 |
| 12/02/19 | 11:13 | CMI CHOWK | 2 | - | - | - | high | high | 94.06 | 25 | 20.08 |
| 12/02/19 | 11:14 | CMI CHOWK | 2 | - | - | - | medium | medium | 51.36 | 24 | 28.71 |
| 12/02/19 | 11:15 | CMI CHOWK | 2 | - | - | - | medium | medium | 56.3 | 19 | 27.42 |
| 12/02/19 | 11:16 | CMI CHOWK | 2 | - | - | - | low | low | 20.99 | 36 | 37.5 |
| 12/02/19 | 11:18 | CMI CHOWK | 2 | - | - | - | medium | medium | 58.67 | 37 | 26.73 |
| 12/02/19 | 11:19 | CMI CHOWK | 2 | - | - | - | medium | medium | 72.46 | 31 | 24.26 |
| 12/02/19 | 11:20 | CMI CHOWK | 2 | - | - | - | medium | medium | 50.45 | 34 | 28.62 |

Copyright © 2019, Dehradun Traffic Monitor. All Rights Reserved.

Figure 5.13: Data access module of prototype

# 6. CONCLUSION AND RECOMMENDATIONS

This study has provided a framework for the estimation of traffic congestion, traffic incident detection and historical trend analysis in near real time from the unstructured data obtained from the inductive loops installed in the study area. The information regarding the traffic is then enriched by integrating data from the social media platform. In present scenario Twitter has the potential to provide a good quantity of data because of its less-strict data privacy norms. In India, so far Twitter is the only social media platform from where quality data can be obtained. So this study mainly focused on the Twitter data for enriching the data obtained from the inductive loops. Natural language processing has been done on the data obtained from the Twitter. And word tokenization and lemmatisation has been performed on the extracted tweets and then tweets were classified into seven different classes. These classes were decided by analysing the major reasons for the traffic in the Dehradun city. The classification was done by using Naïve Bayes Multinomial classification algorithm. The accuracy achieved by this algorithm was 79.15% which is acceptable in the field of words classification (NBM). Also the sentiment analysis was done by using AFINN, in which tweets were labelled as positive, negative and neutral on the basis of sentiments or emotions towards the traffic situation. The NBM classified data will provide the user with the information related to the cause of the traffic while the labelling of the tweets with the sentiments attached to it will help in further improvising the model when integrating it with inductive loop data. Inductive loop data is available in the form of xml format. It includes the information regarding vehicle speed, density of vehicles, road width, vehicle length etc. Among these information provided by inductive loop, speed of vehicle and density of the vehicles have been selected for building the congestion model. Because in the Indian traffic conditions these two parameters have been proved to be the major indicator of traffic. A model proposed by the Maitra et al. (1999)has been utilised in this study for studying the congestion level on the roads. These congestion levels were defined as low, medium and high. Now the information obtained from sentiment analysis (positive, negative and medium) and congestion model (high, low, and medium), both are integrated by applying fuzzy rule method and stored along with the information obtained, regarding the cause of traffic by making use of classification algorithm.

The stored information is then visualised in the web GIS based platform by providing the users with the dashboard application. Traffic monitoring engineers, general public are the major targeted users. This study will effectively provide the user with the information regarding the near real time traffic situation in the city like the congestion level and cause of the traffic and also user can have a historical information about the traffic situation. So, that user can decide to take the necessary action to avoid traffic like can take some different route etc. and also for traffic monitoring engineers, this study will provide the better analysis of near real time traffic so that they can work on improvement of the traffic situation in the city.

## 6.1. Answers to the research question

**Question:** How do the various social media platforms compare with respect to their utility as traffic information source?

*Answer:* On reviewing different literature and exploring the data services provided by the various different social media platforms as discussed in 2.2. It was observed that although Facebook has the largest amount of social users, it lacks behind in providing enough access to the data which could be utilized for traffic related specific applications. Moreover, the users of Facebook have tendency to share more about the personal activities within their social circle. The privacy of the posts also is by default limited to only the acquaintance of the users on the social network which makes availability of public data very less. Other than Facebook, the other popular social networks include Twitter, Instagram, Google + etc. Instagram although has a large user base but is domain specific. It focusses only on the sharing the content by means of images and hence makes itself less common for traffic related posts. Similar to Instagram, Linked IN is also domain specific social network, mostly used for the business purposes or recruitment purpose, making it more specific to a professional social network. Google+ although had seen a large user base in the initial stage but failed to gain popularity and is the least used social network among the above stated social networks with very less posts available on traffic. Lastly comes down Twitter, which is the most favourable social media platform for any kind of data analysis studies. The reason behind Twitter being a popular data source is that the data posted by the users is available publically by default and is easily accessible in both real time stream as well as in for of historic data making use of Twitter developer API(s). Moreover, the community of Twitter is not any domain specific but helps spreading awareness. Finally after manually querying through search feeds of various social media platforms, Twitter has proven to give the best results in terms of road traffic related data. Hence making it the best choice as a social media source for the purpose of road traffic based studies.

**Question:** How to detect spatial extent of non-geotagged traffic related social media feeds?

*Answer:* It was observed that in terms of spatial reference there are two different types of tweets available i.e. geotagged and non-geotagged tweets. Since for congestion estimation, focal point of congestion plays a huge role, it was necessary to determine the spatial extent of all the tweets. The geotagged tweets already had a coordinate associated with them but the non geo tagged tweets had no such information. Hence to make the best use of the non-geotagged tweet, geocoding services were used. For this specific research study, Google Maps Geocoding services were utilized to determine the spatial reference mentioned within the tweets.

**Question:** How to integrate the qualitative data from social media feeds and quantitative data from traffic sensors?

*Answer:* The data available from sensor was quantitative data which was available in semi structured formats, while the data available from social media was in textual format which conveyed some qualitative significance. In order to integrate the two heterogeneous data, the data has to be scaled down to a common scale. For this purpose, it was a practical approach to scale the quantitative data to qualitative. Hence, the sensor data was converted to qualitative form, i.e. congestion levels. While the meaning out of the tweet was extracted by means of sentiment analysis. Once the two qualitative scales are observed, rule based integration systems can be utilized for the integration of the two datasets.

**Question:** Which is the most suitable congestion model for the available dataset?

*Answer:* Many congestion models were reviewed before selection of an appropriate model for the available dataset. Most of the models were based on single input parameters, mainly speed. Although considering just speed was not enough to justify congestion. Hence an empirical model with multiple base inputs proposed by Maitra, Sikdar, & Dhingra (1999) was found to be most suitable model for congestion modelling. The model not only considered multiple parameters, but also assigned them weights according to the proportion of vehicles moving in that direction.

**Question:** How to supplement the missing segments of sensor dataset with the social media feeds?

*Answer:* The sensor data does provide the anomalies in movement patterns, but fails to justify them. This is where twitter comes into play. The integrated data not only provides the possible congestion level at the moment, but also covers the reason behind the anomaly by supplying the type of incident along with the congestion level. Hence the missing segments of the sensor data gets fulfilled with the social media feeds.

**Question:** How to handle and store the integrated data for near real time visualization and historical trend analysis?

*Answer:* In order to handle and store integrated data, the twitter data goes through the process of sentiment analysis, and traffic incident classification and then stored into MongoDB in form of Dictionary key value pairs. While the sensor data is pre-processed, aggregated and then finally passed through congestion model and then finally saved to the MongoDB database. After storage, the integration model then pulls the both the relevant data from the database and the resultant output is appended to the sensor data as new key value pairs.

## 6.2. Recommendations:

Road traffic is a very serious issue especially in developing countries like India. It is causing wastage of time, waste of fuel, environment pollution and mental frustration. The future scope of this study can be recommended as to investigate the method for the further improvement in the quality of information regarding real time traffic condition provided to the users.

- This study utilised only two parameters from the inductive loop data that are speed of the vehicle and density of the vehicles. But there is a scope of including more available parameters like vehicle class, length of the vehicle etc.
- Congestion model can include the information of the natural condition to enrich the model. Information like rainfall, fog, storms, snowfall, hailstorm, landslide etc. have high impact on the traffic condition. Landslide, rainfall and fog are the most important factor in case of the study area and its nearby region. So such information can be beneficial for the user.
- Multi source fusion has proven to give improvised results. In future, data from more sources can be integrated so as to improve the traffic congestion estimation results.

# LIST OF REFERENCES

Abidin, A. F., Kolberg, M., & Hussain, A. (2015). Integrating Twitter Traffic Information with Kalman Filter Models for Public Transportation Vehicle Arrival Time Prediction. In *Big-Data Analytics and Cloud Computing* (pp. 67–82). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-25313-8_5

Adetiloye, T., & Awasthi, A. (2017). Predicting Short-Term Congested Traffic Flow on Urban Motorway Networks. In *Handbook of Neural Computation* (pp. 145–165). Elsevier. https://doi.org/10.1016/B978-0-12-811318-9.00008-9

Adetiloye, T. O. (2018). *Predicting Short Term Traffic Congestion on Urban Motorway Networks*. Montréal, Québec, Canada. Retrieved from https://spectrum.library.concordia.ca/984171/24/Adetiloye_PhD_F2018.pdf

Aichner, T., & Jacob, F. (2015). Measuring the Degree of Corporate Social Media Use. *International Journal of Market Research*, *57*(2), 257–276. https://doi.org/10.2501/IJMR-2015-018

Amin Elsafoury, F. (2013). *MONITORING URBAN TRAFFIC STATUS USING TWITTER MESSAGES*. Retrieved from https://webapps.itc.utwente.nl/librarywww/papers_2013/msc/gfm/elsafoury.pdf

Anand, A., Ramadurai, G., & Vanajakshi, L. (2014). Data Fusion-Based Traffic Density Estimation and Prediction. *Journal of Intelligent Transportation Systems*, *18*(4), 367–378. https://doi.org/10.1080/15472450.2013.806844

Anand, R. A., Vanajakshi, L., & Subramanian, S. C. (2011). Traffic density estimation under heterogeneous traffic conditions using data fusion. In *2011 IEEE Intelligent Vehicles Symposium (IV)* (pp. 31–36). IEEE. https://doi.org/10.1109/IVS.2011.5940397

Aslam, J., Lim, S., & Rus, D. (2012). Congestion-aware Traffic Routing System using sensor data. In *2012 15th International IEEE Conference on Intelligent Transportation Systems* (pp. 1006–1013). IEEE. https://doi.org/10.1109/ITSC.2012.6338663

Asur, S., & Huberman, B. A. (2010). Predicting the Future with Social Media. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (pp. 492–499). IEEE. https://doi.org/10.1109/WI-IAT.2010.63

Awan, M. S. K., & Awais, M. M. (2011). Predicting weather events using fuzzy rule based system. *Applied Soft Computing*, *11*(1), 56–63. https://doi.org/10.1016/j.asoc.2009.10.016

Bachmann, C. (2011). Multi-sensor Data Fusion for Traffic Speed and Travel Time Estimation. Retrieved from https://tspace.library.utoronto.ca/handle/1807/30172

Banaei-Kashani, F., Shahabi, C., & Pan, B. (2011). Discovering patterns in traffic sensor data. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on GeoStreaming - IWGS '11* (pp. 10–16). New York, New York, USA: ACM Press. https://doi.org/10.1145/2064959.2064963

Bird, S., Edward, L., & Ewan, K. (2009). Natural Language Processing with Python. Retrieved March 4, 2019, from https://www.nltk.org/

Camay, S., Brown, L., & Makoid, M. (2012). Role of Social Media in Environmental Review Process of National Environmental Policy Act. *Transportation Research Record: Journal of the Transportation Research Board*, *2307*(1), 99–107. https://doi.org/10.3141/2307-11

Cassidy, M. J., & Bertini, R. L. (1999). Some traffic features at freeway bottlenecks. *Transportation Research Part B: Methodological*, *33*(1), 25–42. https://doi.org/10.1016/S0191-2615(98)00023-X

Cassidy, M. J., & Mauch, M. (2001). An observed traffic pattern in long freeway queues. *Transportation Research Part A: Policy and Practice*, *35*(2), 143–156. Retrieved from https://ideas.repec.org/a/eee/transa/v35y2001i2p143-156.html

Centre for Science and Environment, I. (1989). The environmental problems associated with India's major cities. *Environment and Urbanization*, *1*(1), 7–15. https://doi.org/10.1177/095624788900100102

Chan, R., & Schofer, J. L. (2014). Role of Social Media in Communicating Transit Disruptions. *Transportation Research Record: Journal of the Transportation Research Board*, *2415*(1), 145–151. https://doi.org/10.3141/2415-16

Chu, L., Oh, J.-S., & Recker, W. (2005). Adaptive Kalman filter based freeway travel time estimation.

Coifman, B., Coifman, B., Beymer, D., McLauchlan, P., Malik, J., & B, J. M. (1998). A Real-Time Computer Vision System for Vehicle Tracking and Traffic Surveillance. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.8901

Congress, I. R. (1990). Guidelines for capacity of urban roads in plain areas. *IRC Code of Practice*, *106*.

Daly, E. M., Lecue, F., & Bicer, V. (2013). Westland row why so slow? In *Proceedings of the 2013 international conference on Intelligent user interfaces - IUI '13* (p. 203). New York, New York, USA: ACM Press. https://doi.org/10.1145/2449396.2449423

Davis, N., Joseph, H. R., Raina, G., & Jagannathan, K. (2017). Congestion costs incurred on Indian Roads: A case study for New Delhi. Retrieved from http://arxiv.org/abs/1708.08984

Delhi Economic Survey. (2015). *Economic Survey of Delhi, 2014-15. Delhi Economic Survey*. New Delhi. Retrieved from http://www.indiaenvironmentportal.org.in/files/file/economic survey of Delhi 2014-15.pdf

Desai, D., & Somani, S. (2014). Instinctive Traffic Control and Vehicle Detection Techniques. *International Journal of Scientific & Engineering Research*, *5*(1). Retrieved from http://www.ijser.org

Duan, Z., Liu, L., & Sun, W. (2009). Traffic Congestion Analysis of Shanghai Road Network Based on Floating Car Data. In *International Conference on Transportation Engineering 2009* (pp. 2731–2736). Reston, VA: American Society of Civil Engineers. https://doi.org/10.1061/41039(345)450

Eisenman, S., Fei, X., Zhou, X., & Mahmassani, H. (2006). Number and Location of Sensors for Real-Time Network Traffic Estimation and Prediction: Sensitivity Analysis. *Transportation Research Record: Journal of the Transportation Research Board*, *1964*, 253–259. https://doi.org/10.3141/1964-28

Elmenreich, W. (2002). *Sensor Fusion in Time-Triggered Systems.*

Faouzi, N.-E. El, & Klein, L. A. (2016). Data Fusion for ITS: Techniques and Research Needs. *Transportation Research Procedia*, *15*, 495–512. https://doi.org/10.1016/J.TRPRO.2016.06.042

Fazal, S. (2006). Addressing congestion and transport-related air pollution in Saharanpur, India. https://doi.org/10.1177/0956247806063970

Fei, X., Mahmassani, H. S., & Eisenman, S. M. (2007). Sensor Coverage and Location for Real-Time Traffic Prediction in Large-Scale Networks. *Transportation Research Record: Journal of the Transportation Research Board*, *2039*(1), 1–15. https://doi.org/10.3141/2039-01

Fusco, G., Colombaroni, C., & Sardo, S. (2012). Modeling Road Traffic Congestion by Quasi-Dynamic Traffic Assignment. Retrieved from https://www.semanticscholar.org/paper/Modeling-Road-Traffic-Congestion-by-Quasi-Dynamic-Fusco-Colombaroni/31aea5444e5bab574d60f8dce6cfc916746259f0

Giridhar, P., Amin, M. T., Abdelzaher, T., Kaplan, L., George, J., & Ganti, R. (2014). ClariSense: Clarifying sensor anomalies using social network feeds. In *2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)* (pp. 395–400). IEEE. https://doi.org/10.1109/PerComW.2014.6815239

Grosenick, S. (2012). Real-Time Traffic Prediction Improvement through Semantic Mining of Social Networks. Retrieved from https://digital.lib.washington.edu/researchworks/handle/1773/20911

Guo, J., Xia, J., & Smith, B. L. (2009). Kalman Filter Approach to Speed Estimation Using Single Loop Detector Measurements under Congested Conditions. *Journal of Transportation Engineering*, *135*(12), 927–934. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000071

Hall, D. L., & Llinas, J. (1997). An introduction to multisensor data fusion. *Proceedings of the IEEE*, *85*(1), 6–23. https://doi.org/10.1109/5.554205

He, J., Shen, W., Divakaruni, P., Wynter, L., & Lawrence, R. (2013). Improving Traffic Prediction with Tweet Semantics. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence* (pp. 1387–1393). AAAI Press. Retrieved from http://dl.acm.org/citation.cfm?id=2540128.2540328

He, S. (2012). Analysis Method of Traffic Congestion Degree Based on Spatio-Temporal Simulation. *International Journal of Advanced Computer Science and Applications*, *3*(4). https://doi.org/10.14569/IJACSA.2012.030403

Henry, S., & Koshy, B. I. (2016). *Congestion Modelling for Heterogeneous Traffic.* Kottayam. https://doi.org/10.17577/IJERTV5IS020112

Hofleitner, A., Herring, R., Abbeel, P., & Bayen, A. (2012). Learning the Dynamics of Arterial Traffic From Probe Data Using a Dynamic Bayesian Network. *IEEE Transactions on Intelligent Transportation Systems*, *13*(4), 1679–1693. https://doi.org/10.1109/TITS.2012.2200474

Jiann-Shiou Yang. (2005). Travel time prediction using the GPS test vehicle and Kalman filtering techniques. In *Proceedings of the 2005, American Control Conference, 2005.* (pp. 2128–2133). IEEE. https://doi.org/10.1109/ACC.2005.1470285

JSON. (2019). 18.2. json — JSON encoder and decoder — Python 2.7.16 documentation. Retrieved March 4, 2019, from https://docs.python.org/2/library/json.html

K. Runyoro, A.-A., & Ko, J. (2013). Real-Time Road Traffic Management Using Floating Car Data. *International Journal of Fuzzy Logic and Intelligent Systems*, *13*.

https://doi.org/10.5391/IJFIS.2013.13.4.269

Kaklij, S. P. (2013). *Mining GPS Data for Traffic Congestion Detection and Prediction*. *International Journal of Science and Research* (Vol. 4). Retrieved from www.ijsr.net

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. https://doi.org/https://doi.org/10.1016/j.bushor.2009.09.003

Kiilu, K. K., Okeyo, G., Rimiru, R., & Ogada, K. (2018). Using Naïve Bayes Algorithm in detection of Hate Tweets. *International Journal of Scientific and Research Publications*, *8*(3). https://doi.org/10.29322/IJSRP.8.3.2018.p7517

Kim, S.-S., & Kang, Y.-B. (2007). Congestion Avoidance Algorithm Using Extended Kalman Filter. In *2007 International Conference on Convergence Information Technology (ICCIT 2007)* (pp. 913–918). IEEE. https://doi.org/10.1109/ICCIT.2007.147

Kim, S., & Suh, W. (2014). Modeling Traffic Congestion Using Simulation Software. In *2014 International Conference on Information Science & Applications (ICISA)* (pp. 1–3). IEEE. https://doi.org/10.1109/ICISA.2014.6847430

Kumar, A., & Sebastian, T. (2012). Sentiment Analysis on Twitter. *International Journal of Computer Science Issues*, *9*, 372–378.

Kurkcu, A., Morgul, E. F., & Ozbay, K. (2015). Extended Implementation Method for Virtual Sensors. *Transportation Research Record: Journal of the Transportation Research Board*, *2528*, 27–37. https://doi.org/10.3141/2528-04

Lan, L. W., Sheu, J.-B., & Huang, Y.-S. (2008). Investigation of temporal freeway traffic patterns in reconstructed state spaces. *Transportation Research Part C*, *16*, 116–136. https://doi.org/10.1016/j.trc.2007.06.006

Lécué, F., Tallevi-Diotallevi, S., Hayes, J., Tucker, R., Bicer, V., Sbodio, M. L., & Tommasi, P. (2014). STAR-CITY: semantic traffic analytics and reasoning for CITY. In *Proceedings of the 19th international conference on Intelligent User Interfaces - IUI '14* (pp. 179–188). New York, New York, USA: ACM Press. https://doi.org/10.1145/2557500.2557537

Leduc, G. (2008). Road traffic data: Collection methods and applications. *Working Papers on Energy, Transport and Climate Change*, *1*(55).

Lee, J. H., Gao, S., & Goulias, K. (2015). Can Twitter data be used to validate travel demand models?

Lee, K., Palsetia, D., Narayanan, R., Patwary, M. M. A., Agrawal, A., & Choudhary, A. (2011). Twitter Trending Topic Classification. In *2011 IEEE 11th International Conference on Data Mining Workshops* (pp. 251–258). IEEE. https://doi.org/10.1109/ICDMW.2011.171

Li, H., Caragea, D., Caragea, C., & Herndon, N. (2016). *Disaster Response Aided by Tweet Classification with a Domain Adaptation Approach*. *Special Issue on Human Computer Interaction in Critical Systems* (Vol. 1). Retrieved from http://people.cs.ksu.edu/~ccaragea/papers/jccm17.pdf

Lv, Y., Chen, Y., Zhang, X., Duan, Y., & Li, N. L. (2017). Social media based transportation research: the state of the work and the networking. *IEEE/CAA Journal of Automatica Sinica*, *4*(1), 19–26. https://doi.org/10.1109/JAS.2017.7510316

Maitra, B., Sikdar, P. K., & Dhingra, S. L. (1999). Modeling Congestion on Urban Roads and Assessing Level of Service. *Journal of Transportation Engineering*, *125*(6), 508–514. https://doi.org/10.1061/(ASCE)0733-947X(1999)125:6(508)

Mccallum, A., & Nigam, K. (1998). *A Comparison of Event Models for Naive Bayes Text Classification*. Retrieved from http://www.cs.cmu.edu/~knigam/papers/multinomial-aaaiws98.pdf

Metkari, M., Budhkar, A., & Maurya, A. K. (2013). Development of Simulation Model for Heterogeneous Traffic with no Lane Discipline. *Procedia - Social and Behavioral Sciences*, *104*, 360–369. https://doi.org/10.1016/J.SBSPRO.2013.11.129

MongoDB. (2008). PyMongo 3.7.2. Retrieved March 4, 2019, from https://api.mongodb.com/python/current/

Necula, E. (2014). Dynamic Traffic Flow Prediction Based on GPS Data. In *2014 IEEE 26th International Conference on Tools with Artificial Intelligence* (pp. 922–929). IEEE. https://doi.org/10.1109/ICTAI.2014.140

Necula, E. (2015). Analyzing Traffic Patterns on Street Segments Based on GPS Data Using R. *Transportation Research Procedia*, *10*, 276–285. https://doi.org/10.1016/J.TRPRO.2015.09.077

Nellore, K., & Hancke, G. P. (2016, January 27). A survey on urban traffic management system using wireless sensor networks. *Sensors (Switzerland)*. Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/s16020157

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. Retrieved from http://arxiv.org/abs/1103.2903

Numpy. (2018). NumPy — NumPy. Retrieved March 4, 2019, from http://www.numpy.org/

Oh, J.-S., Oh, C., Ritchie, S. G., & Chang, M. (2005). Real-Time Estimation of Accident Likelihood for Safety Enhancement. *Journal of Transportation Engineering*, *131*(5), 358–363. https://doi.org/10.1061/(ASCE)0733-947X(2005)131:5(358)

Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *LREC*. Retrieved from https://www.semanticscholar.org/paper/Twitter-as-a-Corpus-for-Sentiment-Analysis-and-Pak-Paroubek/ad8a7f620a57478ff70045f97abc7aec9687ccbd

Pamuła, T., & Król, A. (2016). The Traffic Flow Prediction Using Bayesian and Neural Networks (pp. 105–126). Springer, Cham. https://doi.org/10.1007/978-3-319-19150-8_4

Pan, B., Zheng, Y., Wilkie, D., & Shahabi, C. (2013). Crowd sensing of traffic anomalies based on human mobility and social media. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL'13* (pp. 344–353). New York, New York, USA: ACM Press. https://doi.org/10.1145/2525314.2525343

Pandas. (2019). Python Data Analysis Library — pandas: Python Data Analysis Library. Retrieved March 4, 2019, from https://pandas.pydata.org/

Patel, J., & Gundaliya, P. J. (2016). Estimation of Level of Service through Congestion-A Case Study of Ahmedabad City. *International Research Journal of Engineering and Technology*. Retrieved from www.irjet.net

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in {Python}. *Journal of Machine Learning Research*, *12*, 2825--2830. Retrieved from https://scikit-learn.org/stable/about.html#citing-scikit-learn

Petalas, Y. G., Ammari, A., Georgakis, P., & Nwagboso, C. (2017). A Big Data Architecture for Traffic Forecasting Using Multi-Source Information (pp. 65–83). Springer, Cham. https://doi.org/10.1007/978-3-319-57045-7_5

Sasaki, K., Nagano, S., Ueno, K., & Cho, K. (2012). Feasibility Study on Detection of Transportation Information Exploiting Twitter as a Sensor. Retrieved from https://www.semanticscholar.org/paper/Feasibility-Study-on-Detection-of-Transportation-as-Sasaki-Nagano/717443b82dfe384b4cbe7072e5bec67b28d107a1

Schönhof, M., & Helbing, D. (2007). Empirical Features of Congested Traffic States and Their Implications for Traffic Modeling. *Transportation Science*, *41*(2), 135–166. https://doi.org/10.1287/trsc.1070.0192

Schulz, A., Ristoski, P., & Paulheim, H. (2013). I See a Car Crash: Real-Time Detection of Small Scale Incidents in Microblogs (pp. 22–33). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-41242-4_3

Shen, W., & Zhang, H. M. (2009). On the morning commute problem in a corridor network with multiple bottlenecks: Its system-optimal traffic flow patterns and the realizing tolling scheme. *Transportation Research Part B: Methodological*, *43*(3), 267–284. https://doi.org/10.1016/J.TRB.2008.07.004

Stambaugh, C. L. (2013). Social Media and Primary Commercial Service Airports. *Transportation Research Record: Journal of the Transportation Research Board*, *2325*(1), 76–86. https://doi.org/10.3141/2325-08

Statista. (2018). • Statista - The Statistics Portal for Market Data, Market Research and Market Studies. Retrieved January 28, 2019, from https://www.statista.com/

Sun, Z., Guo, M., Liu, W., Feng, J., & Hu, J. (2009). Multisource Traffic Data Fusion with Entropy Based Method. In *2009 International Conference on Artificial Intelligence and Computational Intelligence* (pp. 506–509). IEEE. https://doi.org/10.1109/AICI.2009.392

Teodorovic, D., Lucic, P., Popovic, J., Kikuchi, S., & Stanic, B. (2001). Intelligent isolated intersection. *Fuzzy Systems, 2001. The 10th IEEE International Conference.*, *1*, 276–279. Retrieved from https://ieeexplore.ieee.org/abstract/document/1007302/

Thant Lwin, H., & Thu Naing, T. (2015). IJARCCE Estimation of Road Traffic Congestion using GPS Data. *International Journal of Advanced Research in Computer and Communication Engineering*, *4*. https://doi.org/10.17148/IJARCCE.2015.41201

Thianniwet, T., Phosaard, S., & Pattara-Atikom, W. (2010). Classification of Road Traffic Congestion Levels from Vehicle's Moving Patterns: A Comparison Between Artificial Neural Network and Decision Tree Algorithm (pp. 261–271). https://doi.org/10.1007/978-90-481-8776-8_23

Treiber, M., & Kesting, A. (2010). Calibration and validation of models describing the spatiotemporal evolution of congested traffic patterns. Retrieved from http://arxiv.org/abs/1008.1639

Tso, B., & Mather, P. M. (2009). *Classification methods for remotely sensed data*. CRC Press. Retrieved from https://www.crcpress.com/Classification-Methods-for-Remotely-Sensed-Data-Second-Edition/Mather-Tso/p/book/9781420090727

Twitter. (2011). Twitter Streaming API. Retrieved February 12, 2019, from https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data.html

Uttarakhand Police. (1988). Traffic Violations and Fine: Traffic Violations and Fine. Retrieved February 27, 2019, from https://uttarakhandpolice.uk.gov.in/pages/display/143-traffic-violations-and-fine

Vinsel Lee. (2016). The Man Who Invented Intelligent Traffic Control a Century Too Early - IEEE Spectrum. Retrieved August 24, 2018, from https://spectrum.ieee.org/tech-history/dawn-of-electronics/the-man-who-invented-intelligent-traffic-control-a-century-too-early

Vlahogianni, E., Karlaftis, M., & Golias, J. (2014). Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C: Emerging Technologies*, *43*.

Wang, S., Li, F., Stenneth, L., & Yu, P. S. (2016). Enhancing Traffic Congestion Estimation with Social Media by Coupled Hidden Markov Model. In *Machine Learning and Knowledge Discovery in Databases* (pp. 247–264). Springer, Cham. https://doi.org/10.1007/978-3-319-46227-1_16

Wang, S., Zhang, X., Cao, J., He, L., Stenneth, L., Yu, P. S., … Huang, Z. (2017). Computing Urban Traffic Congestions by Incorporating Sparse GPS Probe Data and Social Media Data. *ACM Transactions on Information Systems*, *35*(4), 1–30. https://doi.org/10.1145/3057281

Wang, X., Peng, L., Chi, T., Li, M., Yao, X., & Shao, J. (2015). A Hidden Markov Model for Urban-Scale Traffic Estimation Using Floating Car Data. *PLOS ONE*, *10*(12), e0145348. https://doi.org/10.1371/journal.pone.0145348

Wang, Y., & L. Nihan, N. (2005). Freeway traffic speed estimation using single loop outputs, *1*(Vehicle Information and Communication Systems).

White, J., Thompson, C., Turner, H., Dougherty, B., & Schmidt, D. C. (2011). WreckWatch: Automatic Traffic Accident Detection and Notification with Smartphones. *Mobile Networks and Applications*, *16*(3), 285–303. https://doi.org/10.1007/s11036-011-0304-8

Xiang, Z., & Gretzel, U. (2010). Role of social media in online travel information search. *Tourism Management*, *31*(2), 179–188. https://doi.org/10.1016/J.TOURMAN.2009.02.016

Xu, S., Li, S., & Wen, R. (2018). Sensing and detecting traffic events using geosocial media data: A review. *Computers, Environment and Urban Systems*. https://doi.org/10.1016/J.COMPENVURBSYS.2018.06.006

Zhang, D., Nagurney, A., & Wu, J. (2001). On the equivalence between stationary link flow patterns and traffic network equilibrium. *Transportation Research Part B: Methodological*, *35*(8), 731–748. Retrieved from https://ideas.repec.org/a/eee/transb/v35y2001i8p731-748.html

Zhang, S., Tang, J., Wang, H., & Wang, Y. (2015). Enhancing Traffic Incident Detection by Using Spatial Point Pattern Analysis on Social Media. *Transportation Research Record: Journal of the Transportation Research Board*, *2528*, 69–77. https://doi.org/10.3141/2528-08

Zheng, W., Lee, D.-H., & Shi, Q. (2006). Short-Term Freeway Traffic Flow Prediction: Bayesian Combined Neural Network Approach. *Journal of Transportation Engineering*, *132*(2), 114–121. https://doi.org/10.1061/(ASCE)0733-947X(2006)132:2(114)

Zheng, Z., Wang, C., Wang, P., Xiong, Y., Zhang, F., & Lv, Y. (2018). Framework for fusing traffic information from social and physical transportation data. *PLOS ONE*, *13*(8), e0201531. https://doi.org/10.1371/journal.pone.0201531

Zhou, X., & List, G. F. (2010). An Information-Theoretic Sensor Location Model for Traffic Origin-Destination Demand Estimation Applications. *Transportation Science*, *44*(2), 254–273. https://doi.org/10.1287/trsc.1100.0319