2016 Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), Sept. 21-24, 2016, Jaipur, India

Vehicular Traffic Analysis From Social Media Data

Himanshu Shekhar*, Shankar Setty†, Uma Mudenagudi‡
B.V. Bhoomaraddi College of Engineering and Technology, Hubli, India
Email: *himanshu508@hotmail.com, †shankar@bvb.edu, ‡uma@bvb.edu

***Abstract*—In this paper, we address the problem of vehicular traffic congestion occurring in densely populated cities. Towards this we propose to provide a framework for optimal vehicular traffic solution using social media live data. Typically, the traffic congestion problem addressed in literature focuses on usage of dedicated traffic sensors and satellite information which is quite expensive. However, many urban commuters tend to post updates about traffic on various social media in the form of tweets or Facebook posts. With the copious amount of data made available upon traffic problems on social media sites, we collect historical data about traffic posts from specific cities and build a sentiment classifier to monitor commuters' emotions round the clock. The knowledge is used to analyze and predict traffic patterns in a given location. Also we identify the probable cause of a traffic congestion in a particular area by analyzing the collected historical data. Through our work, we are able to present an uncensored, economical and alternative approach to traditional methods for monitoring traffic congestion.**

Keywords—Social Media, Vehicular Traffic, Sentiment Analysis

I. Introduction

Urban density is continuously plagued by population growth. Almost all the commuters are affected by traffic congestion, regardless of whether they are contributing to it by driving a vehicle or trying to avoid it by using mass transit. Despite the trend of authorities to enforce and straighten the public transportation infrastructure in high density cities, traffic congestion can be expected to worsen[1]. However, from the viewpoint of individual drivers, we can assume that they are rational decision makers who will do their best to minimize the trip time provided they have adequate information[2]. This presents an opportunity to reduce traffic congestion by providing accurate traffic predictions to drivers as they can each optimize their individual travel plans. Individuals directly realize the benefit of this approach through shorter trip times. Simultaneously, the entire transportation system benefits from fewer vehicles unknowingly heading into areas of congestion, exacerbating backups in problem spots. While there are clear benefits to providing this information to drivers, it is not abundantly clear how to formulate such traffic predictions.

The costs of traffic congestion in India is difficult to quantify but in the year 2012, it was estimated to account for the loss of 600 billion Indian Rupees, including fuel wastage and productivity delays[3]. Traffic congestion increases the time required to traverse road segments, so in addition to increased fuel consumption, it results in increased air pollution[4] and health problems such as heart attacks[5], premature birth of infants[6] and psychological stress[7].

The prediction in the behavior of complex systems, such as accurate traffic speeds in regional transportation systems has been explored extensively in literature[2][8]. However, any non-recurring or unplanned event, such as traffic accident, or even inadvertent climatic condition disrupts prediction accuracy. Yet taking into account these events is very important, as these are precise events about which the commuters must be aware of whenever they are undertaking any trip. There are many active fields of research to autonomously detect such events, for example automated incident detection from traffic flow data and computer vision. However, these methods are still in development and do not produce an entirely accurate results when compared to actual human observations. This raises the need to introduce new techniques to identify and predict the traffic congestion without any human input. One possible solution of this scenario is to use the data extracted from social network to provide a new system for traffic congestion analysis.

Mostly traffic analysis is performed using traditional data sources, which have the the advantage of being reliable, well-governed, and of standard quality. However, these sources are plagued with their own practical challenges. It's very difficult to maintain and provide continuous updation to such a data source that may vary greatly in respect of volume, reliability and streaming frequency. On the other hand, non-traditional sources such as streaming social media information usually have low reliability, negligible amount of governance and eclectic size. Still, if properly handled, such user-generated data sources can be easily used to augment the conventional information sources. For our research work, we have focused primarily on Facebook and Twitter, two of the most popular social media platforms which allow users to share their updates globally entirely uncensored. Additionally, most of these updates are geo-tagged which helps in determining the origin location of the posts. The updates can be posted through a host of front-end clients, such as the web, dedicated mobile apps, email, text messaging and many other third party clients. The data thus generated is exhaustive in size and almost entirely user-centric. However, due to diverse and colloquial nature of posts from users' located in different geographical location, the determination of general emotion and behavior expressed by them remains a particularly hard problem to solve.

The primary motivating factor for our work is the analysis of real-time traffic conditions in cities or locations as a task which according to our knowledge has not been automated and is mostly carried out by the manual perusal of historic data and real-time imaging by satellites. The live satellite traffic information does not provide any cause for the traffic congestion. We propose to automate traffic congestion analysis by mining the knowledge from the data present in social media. Facebook feeds and tweets from Twitter can be used to figure out frequent congestion points due to specific scenarios (viz. sporting event, faulty signal, maintenance work), which would be a valuable insight to eliminate the problem at that particular location. In our work, we demonstrate vehicular traffic congestion by analyzing the effects of traffic congestion on affected people by analyzing the user-generated social media posts for the same. The behavior and emotions of affected users are categorized and clustered according to various locations. The data is gathered over time and trained which gives the probable state of traffic congestion in a particular location at a given time slot. Over time, if the particular congestion is found consistent in a particular location, we employ Natural Language Processing (NLP) technique in order to detect the root cause of the problem.

Towards this, the major contributions of our work are:

- Analysis of traffic problems for specific locations for a given time window.
- Identification of the cause of a recurring congestion.

II. Related Work

In this section we discuss some of the works related to our approach. First we analyze the usage of NLP in various scenarios, and then we look over various traffic analysis techniques. Finally we examine the traffic analysis work involving NLP.

A. Natural language processing in social media

NLP in social media has always been an interesting topic in the research community. Social media is easily recognized as an immense source of uncensored data. The basic framework of analysing NLP was tried in [9] where authors made a black box for efficient text processing of streaming online social networks data. The work was further extended by Badlwin *et al.*[10] in surveying the presence of noise in social media texts. It is understandable that extracting and analyzing such immense data source is full of numerous challenges and constraints, such were evaluated by Maynard *et al.*[11].

The usage of data extracted from social media delves into many fields. Some notable examples comprise the work by Yang *et al.*[12] wherein knowledge from social media was used for Drug Safety. In their work, authors used association mining and proportional reporting ratios to mine the associations between drugs and adverse reactions from the user contributed content in social media. Other wide-ranging examples include mining of social media for the purpose of polling on user-created subjects[13], analysing the effects of natural disasters by noticing the trends and user sentiments from tweets[14][15], and proposal of the usage of social media posts during a mass emergency[16][17]. Social media posts were also used to determine the general emotional condition of any particular user. Yang *et al.*[18] built an online happiness-profile of individual Twitter users by performing a sentiment analysis upon their tweets ranging over a period of two years. Numerous other examples are present in literature, where NLP in social media have been used to extract valuable information about topics of interest.

B. Traffic analysis

Traffic congestion prediction is a very popular subject in computational research. The work in this field ranges from prediction-based guidance taking in account of commuter's choices during travels[19] to building a model by detecting spatio-temporal correlations in a road network[20]. Traffic prediction has many application in real-time, such as a creation of a real-time trip information system for a large GPS-enabled Singaporean taxi company by Balan *et al.*[8]. Short term traffic forecasting was achieved by Vlahogianni *et al.* by making use of complex mathematical traffic models[21]. We noticed that most of current traffic prediction technologies make use of GPS derived satellite data for the purpose of detecting traffic congestion[22].

C. Traffic analysis using NLP

Prediction of traffic using social media tools is relatively new, and has been on advance particularly with the increased usage of Web 2.0. In [23], authors demonstrate a monitoring tool which indicates an increase in traffic flow for particular region by observing the increase in rate of tweets in that region. In [24], authors proposed short-term traffic optimization techniques through social media posts for traffic analysis. We also observe that Septianusa *et al.*[25] provided a case study by mining the data from twitter posts to monitor the traffic conditions going in and out of several lanes from the port of Tanjung Priok of Indonesia. Particularly, Khatri[26] developed a model for the congestion analysis of traffic in United States by means of tweets. Khatri proposed a model for the perception analysis of the tweets by first isolating the congestion and incident related tweets and then quantifying the perception indicated in tweets through the means of NLP and Artificial Intelligence methods. Further, semantic web model styled as STAR CITY was used by researchers to detect, diagnose and predict the traffic congestion problems in the cities of Dublin, Ireland and Bologna, Italy[27]. The comparisons and lessons from implementing STAR CITY in metro cities were discussed by the researchers in [28]. In similar manner, Grosenick [29] demonstrated in his thesis a method of traffic prediction through a combination of sensors placed along the highways and the tweets captured pertaining to traffic in that region. This method, though highly reliable, is expensive as it requires the usage of physical sensors to provide the results. We aim to overcome this problem by making a system entirely upon social media posts.

Drawing inspiration from the works of [26], [25] and [27] we propose a process of detecting traffic congestion and its cause in selected metro cities in India through the means of social media data.

III. Real-time traffic scenario for analysis

The basic underlying principle behind our system is the assumption that there is a constant stream of social media posts being generated on the web which can be easily extracted and mined. There is usually a spike in the quantity of such posts with the occurrence of any event with may affect the masses, such as a sporting event, festival, rally etc. Usually the spike in streaming rate for a particular location indicates a gathering of people in the area under observation, which in turn indicates a possibility of occurrence of a traffic congestion. Also, many metro cities maintain a dedicated social media page for traffic grievances, which serve as a portal for commuters to share their experience.

For our application, we acquire data from Facebook and Twitter based on the users post pertaining to traffic conditions in different locations. We then eliminate noisy data unrelated to traffic and extract features like location, time-stamp etc. from the feeds. Sentiment analysis is done to perceive the general mood of users at a specific time of the day. Based on these insights, a decision tree is constructed to display traffic sensitive optimized routes. The results of these data set is stored and analyzed over a period of time. If a recurring pattern of congestion is found, we use a keyword extraction algorithm to determine which factors are more prevalent in causing such a pattern.

Official traffic broadcasts do not suffer from the problems inherent in generalized mining of social media data. Major metropolitan transit authorities, such as the Bangalore Traffic Police, Delhi Traffic Police, Brihanmumbai Electric Supply and Transport (BEST - Mumbai Traffic Authority) broadcast messages on their social media accounts. The trained staff of a regional transportation authority is intentionally monitoring and reporting on traffic conditions affecting the majority of their transportation system. It can be assumed that these primary sources cover the majority of interesting roadways and use consistent conventions for naming events, locations, and severity. These official accounts are also used by users to post their own messages and grievances, thus providing a valuable amount of data. The data thus collected consists of a huge assortment of wildly diverse emotions from a variety of users from selected cities around India. Such a dataset presents a number of challenges, few of them noted as follows:

- **Inscrutable sentences:** There is an ardent lack of usage of proper English grammar in social media post. Also, most of these post tend to be use colloquial language. In addition to this, slang terms and *internet lingo* such as LOL for "Laughing Out Loud" are abundantly present in the posts. Traditional NLP tools are usually unable to parse and analyze such data.
- **Size of posts:** The message length in the posts varies greatly, thus creating a scenario of either *too many* or *too few* words for sentiment analyzer.
- **Presence of local locations:** Posts usually refer to colloquial references [17], such words are not mentioned in any dictionary. Identification of such locations pose a major problem to the parser.
- **Presence of noise:** Social network data is extremely noisy, and its possible to capture redundant posts. Extensive pre-processing is required to tackle this issue.

IV. Vehicular traffic analysis framework

In this section, we address the problem of traffic congestion occurring in densely populated cities by proposing a framework for optimal vehicular traffic solution using social media data. The proposed framework comprises of extraction of data, analysis of data and identification of congestion causing events. Fig. 1 shows an overview of the proposed framework for traffic congestion analysis.

A. Data extraction

The source of data for our work are from Twitter and Facebook. Tweets are collected via Twitter streaming API, wherein traffic related tweets are filtered by applying a series of weighted keywords. Obtained tweets are categorized on the basis of their origin by applying geo-filter tag on GPS enabled tweets as well as keyword match. In case of Facebook, posts are extracted from dedicated traffic information pages which are crawled via Facebook Graph API. Data from tweets and Facebook pages originating from different metro cities in India like Bangalore, New Delhi, Kolkata, Chennai and Mumbai are maintained.

1) Redundancy check: The posts may be redundant due to re-tweets, sharing of posts by users, and recapturing of older data. This issue is prevented by the usage of a regular expression method which randomly checks sections of incoming posts against the exiting ones for similar string patterns. If there is a higher percentage of pattern-match, the captured post is summarily dropped.

Detection of re-tweets remains a difficult problem to solve. However, its importance in modeling traffic data from secondary to the proper aggregation of primary data sources, such as initial tweets from transit authorities. While the accurate capture and aggregation of traffic broadcast re-tweets could provide incremental value in traffic predictions, it is theorized that this is not critical in prediction accuracy. Although number of retweets emphasize in the severity of a particular congestion event, they might be posted by users on a much later time from the actual happening of the event and thus would not be particularly relevant during congestion prediction. Instead of actually storing all the retweets, its much more feasible to store the count of retweets on the original post.

2) Storage of Data: The posts which pass the redundancy check are systematically stored. The stored data consists of the posts' message content, the username of the person who posted the message and time-stamp of the message. Also additional information such as number of re-tweets and comments upon the Facebook posts are stored in order to determine the severity of a event causing traffic congestion.

B. Analysis of data

The categorized data is passed on for analysis and interpretation which consists of methods such as check the data for geo-tagging and location identification, user emotion analysis and optimized route generation.

1) Location identification: The location of originating post is the foremost matter of interest in any type of traffic congestion analysis. To achieve this, we apply a geo-filter tag on the posts which returns the point of posts' origin. Even so, we observe that in most cases, the posts are from the people who are not physically present at the location which is being mentioned in the post. Most often, in such a case the traffic congestion location is mentioned within the post. To solve this issue, we initially identify all the prepositions present in the sentence (eg. *in, at, on* etc.). This is so because in common sentence formations, a location is almost always preceded by a preposition (eg. *Traffic Jam at Church Road.*. Once the prepositions are identified, the words immediately

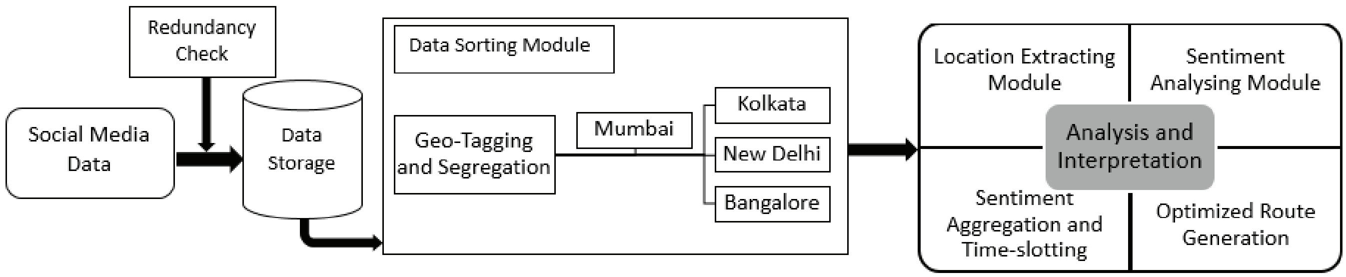


Fig. 1: An overview of proposed vehicular traffic analysis framework

succeeding it are sent to Google Maps API. The API returns the geographical co-ordinates if the inquiry is within the perimeter of the city. We then store this corresponding location and time-stamp along with the posts. We also track the isolated location names mentioned in a particular post.

2) User's emotion analysis: Each categorized post is fed to a sentiment analysis method. A word dictionary database consisting of weighted sentiment rating for every word is created. The weighed dictionary being used for this purpose is acquired from SentiStrength online tool [30] which ranks each word with a ranking providing the level of positivity or negativity of that word [31]. We feed each post to a method which splits it into words and check the sentiment associated with each particular word. A word having a positive sentiment rating is given a positive integer value and vice-versa, wherein a greater positive value indicates the word is having a higher level of positivism and a greater negative value indicates that the word has higher level of negativism. Abbreviations or slangs are used against a special slang dictionary which contains a weighed list of most of internet slangs. The presence of negation words, such as *don't, not etc.* which might change the overall sentiment of the post is handled through a special *negating words list* which reverses the overall sentiment appropriately. For example, *I like you* will be given a sentiment rating of 3 but *I don't like you* will be given a rating of -3. We observe that most of upper and lower limit of our sentiment rating for posts fall in the range of -10 to 5. This information coupled with the mentioned location name and the time-stamp of the post gives different level of negativity or positivity present in posts and helps to determine the level of traffic congestion present in a particular area.

3) Optimized route generation: The sentiment generated in each post is associated with the location. Table I shows categorization of average user's sentiment for a particular location within a city based on hourly time-slots. This data is supplemented with inputs from a short-term calendar which contains the detail of any special event happening in the city which may cause a traffic congestion in a given location. The final trained dataset is fed to Google Maps API with instructions to progressively avoid those locations between source and destination which might have the users sentiment as negative in the given time slot.

C. Traffic congestion cause analysis

In our work, we detect the possible reason for traffic congestion in particular area by analyzing the posts pertaining to the area. Events can be expressed by text elements such as verbal predicates and their arguments (*"traffic signal not working"*), noun phrases headed by nominalizations (*"private bus parking next to traffic light"*) and event-referring nouns (*"slow moving traffic"*, *"road accident"*).

We realize that traffic scenarios are frequently changing, and events like *"road accident"* or *"parked bus"* might not be relevant to commuters after a few hours of their conception. For this reason, we limit our keyword search parameter to posts with a upload timestamp of not more than three hours before the query is posed by the user.

The user may search the cause of a congestion at a particular time if interested, or may give a location name and look for all possible historic reasons of congestion at a particular place. In the former case the three hour window is applied on the keyword search result, while in the latter case the search algorithm looks for all possible matches for the queried location. Once the common patterns are identified, they are broken down into separate words, and are sorted on the basis of their count. This ensures that the problem mentioned repeatedly by various users are identified as the major problem which is the most probable cause of traffic disruption in a particular location or area.

V. Results and Discussion

We collected dataset from Facebook and Twitter on five cities of India, viz. Bangalore, New Delhi, Mumbai, Kolkata and Chennai over a period of three months (January 2016 - March 2016). The results demonstrated in this paper are based on the analysis performed on the cities of Bangalore and New Delhi. The statistics of the data set from Bangalore and New Delhi are given as follows:

- Tweets collected for Bangalore: 2163
- Facebook posts collected for Bangalore: 1458
- Tweets collected for New Delhi: 1856
- Facebook posts collected for for New Delhi: 987
- Average length of tweets: 21 words (108 characters) having a standard deviation of 7.2
- Average Facebook Post length: 47 words (97 characters) having a standard deviation of 11.8

The results discussed in Table I demonstrate the sentiment analysis and its interpretation for various location in the city.

TABLE I: Average user sentiment analysis w.r.t. particular location based on hourly time-slots. *Date: May 12th, 2016*

Streets of Bangalore	Time	Avg. User Sentiment	Conclusion
Church Street	6pm-7pm	-2	Moderate Congestion
Church Street	8pm-9pm	-7	Severe Congestion
Adugodi road	1pm-2pm	1	No Congestion
MG Road	7pm-8pm	-6	Severe Congestion
Jainagar Road	8am-9am	-2	Moderate Congestion
Streets of New Delhi	**Time**	**Avg. User Sentiment**	**Conclusion**
Hauz Khas	7pm-8pm	-7	Severe Congestion
Rajiv Chowk	11am-12pm	-1	Moderate Congestion
Preet Vihar	4pm-5pm	3	No Congestion
Lodi Gardens	8pm-9pm	-1	Moderate Conegstion
Janpath	7pm-8pm	2	No Congestion

The severity of a traffic condition is determined by the average user sentiment value for that location. Mostly, a positive sentiment is interpreted as having no traffic congestion, whereas sentiment ranging from 0 to -5 is considered to indicate moderate traffic congestion and -5 or less is considered to denote severe traffic congestion. This scale can be perfected over time by analyzing the results and comparing with commercial traffic notification services for reference.

Further, we observe the causes of a traffic congestion as suggested by the proposed framework. Table II shows the result queried on MG Road in city of Bangalore and Connaught Place in New Delhi, both being one of the busiest areas of their respective cities. By analyzing the results presented in Table II, we can interpret that on a given day, there is usually a traffic congestion at MG Road in the evening which is mostly caused by high volume of traffic during that period. Afternoon (1pm - 2pm) is usually congestion free, which causes our method to display general keywords from the posts of that period. However a historic traffic congestion search on the same location yields interesting results, giving information that traffic (high volume of traffic), cricket event (owing to the proximity of the cricket stadium in that region) and cinema event are the major cause of traffic woes in that area. Similarly in New Delhi, Connaught place, a popular area with tourists, suffers most of its traffic problems due to high volume of office-goers, tourist congregation and illegal parking. Volume of social media posts pertaining to a specific event causing traffic congestion for the time period of January to March 2016 has been demonstrated in Fig. 2. Such an insight is valuable both from a user's perspective who wants to visit the locality and from authorities perspective, who would intend to ease the traffic load in the region.

In retrospect, we compare the results from our system with those provided by commercially available GPS tools. For this purpose, we obtain traffic status of various localities of Bangalore city through HERE Drive+ application and check this information against our system. A sample of the results are provided in Fig. 3. We note the results provided by both the systems are very consistent, with additional information about possible cause of congestion is provided by our system.

We consider the event recognition as an work in progress, but the initial results have shown to be very promising. Our proposed framework is cost effective when compared to conventional means. We have not applied or framework to the Chennai city data set, where there is a lot of colloquial language usage which cannot be parsed by our sentiment analysis or pattern recognition algorithm as it is designed solely for English language. In such a scenario, the quality of result is expected to lower.

There is an inadvertent biasing in results as people with limited access to digital technology are not represented and thus the result excludes the data for this population. Human factors are highly accountable while handling the social media data, and the generated result can be assumed to be correct. Further, since we have limited the approach to only the metro cities of India, there is an increase in accuracy as the metropolitan population is more exposed to social media than of small towns and cities. We have also introduced a method to detect the cause of the traffic problem in a particular location, which as per our knowledge is not present even in the most advanced traffic prediction system commercially available. Our system can be extended to be combined with the traffic detection system through satellites and sensors[22]. Doing so will give a complete live stream traffic prediction system which would encompass human-factor and *word-of-mouth* knowledge from the commuters to a user, including the possible cause of a traffic congestion. This system is particularly useful in checking for traffic condition when internet connectivity is not available, as the possible *word-of-mouth* knowledge about usual traffic condition in a particular place can be extracted from the historical data stored in the application. Even though if the user has access to commercially available maps and traffic data, its worthwhile to know the views of commuters and find the possible cause of congestion at a particular location.

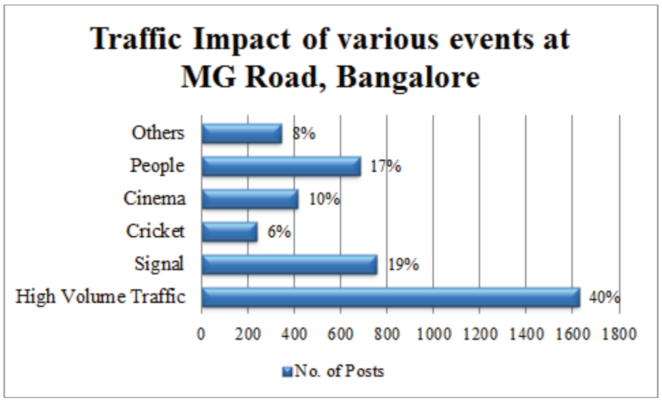
VI. Conclusions and Future Work

In this work we have successfully presented the usage of social media posts as an alternative approach to analyze the traffic congestion and its causes in selected cities. We have demonstrated the results on two Indian cities and are working on to expand it to more. The development of our traffic congestion analysis framework solely on the basis of social media posts is very much work in progress, the initial results are very much promising and we are confident that the strategies implemented by us would lead to a complete system successfully.

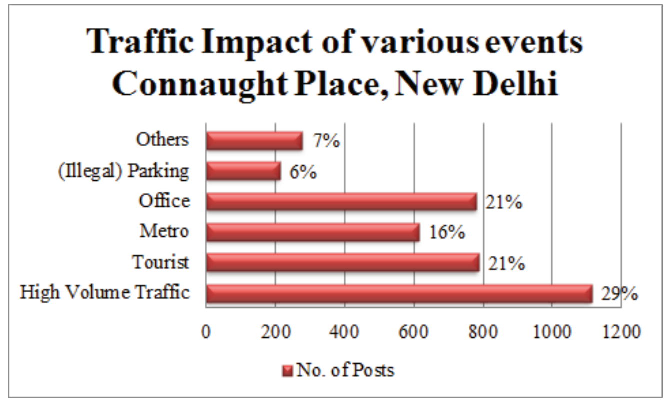
There are many improvements which can be made in this system in terms of using further linguistic and contextual clues. The application described here is an initial step to a more complete system, and it also contextualizes the work within a wider framework of social media monitoring.

TABLE II: Probable cause of traffic congestion at a local place. Event keywords sorted with respect to frequency of occurrence. *Date: May 12th, 2016*

Street of Bangalore	Time	Avg. User Sentiment	Event Keywords
MG Road	6pm-7pm	-1	people, high, traffic
MG Road	8pm-9pm	-6	traffic, signal, rally
MG Road	10am-11am	1	people, metro
MG Road	No time specified, historic data search	-1	traffic, cricket, cinema, people
Street of New Delhi	**Time**	**Avg. User Sentiment**	**Event Keywords**
Connaught Place	7pm-8pm	-7	traffic, metro, office
Connaught Place	11am-12pm	-1	shoppers, tourist, rally, vendors
Connaught Place	4pm-5pm	-2	tourist, metro, parking, illegal
Connaught Place	no time specified, historic data search	-3	tourist, rally, metro, office



(a)



(b)

Fig. 2: Analysis of impact percentage due to various events on *(a)* MG Road, Bangalore, and *(b)* Connaught Place, New Delhi over historical social media data.

REFERENCES

[1] S. Melia, G. Parkhurst, and H. Barton, "The paradox of intensification," *Transport Policy*, vol. 18, no. 1, pp. 46–52, 2011.

[2] J. Van Lint, "Reliable real-time framework for short-term freeway travel time prediction," *Journal of transportation engineering*, vol. 132, no. 12, pp. 921–932, 2006.

[3] D. K. Dash, "India loses Rs 60,000 crore due to traffic congestion: Study," http://timesofindia.indiatimes.com/india/India-loses-Rs-60000-crore-due-to-traffic-congestion-Study/articleshow/13678560.cms, 2012.

[4] J. Shawe-Taylor, T. De Bie, and N. Cristianini, "Data mining, data fusion and information management," in *Intelligent Transport Systems, IEE Proceedings*, vol. 153, no. 3. IET, 2006, pp. 221–229.

[5] A. Peters, S. von Klot, M. Heier, C. H. R. in the Region of Augsburg Study Group *et al.*, "Exposure to traffic and the onset of myocardial infarction," *ACC Current Journal Review*, vol. 14, no. 1, p. 19, 2005.

[6] J. Currie and W. R. Walker, "Traffic congestion and infant health: Evidence from e-zpass," National Bureau of Economic Research, Tech. Rep., 2009.

[7] D. Stokols, R. W. Novaco, J. Stokols, and J. Campbell, "Traffic congestion, type a behavior, and stress." *Journal of Applied Psychology*, vol. 63, no. 4, p. 467, 1978.

[8] R. K. Balan, K. X. Nguyen, and L. Jiang, "Real-time trip information service for a large taxi fleet," in *Proceedings of the 9th international conference on Mobile systems, applications, and services*. ACM, 2011, pp. 99–112.

[9] D. Preotiuc-Pietro, S. Samangooei, T. Cohn, N. Gibbins, and M. Niranjan, "Trendminer: An architecture for real time analysis of social media text," in *Proceedings of the workshop on real-time analysis and mining of social streams*, 2012.

[10] T. Baldwin, P. Cook, M. Lui, A. MacKinlay, and L. Wang, "How noisy social media text, how diffrnt social media sources?" in *IJCNLP*, 2013, pp. 356–364.

[11] D. Maynard, K. Bontcheva, and D. Rout, "Challenges in developing opinion mining tools for social media," *Proceedings of the@ NLP can u tag# usergeneratedcontent*, pp. 15–22, 2012.

[12] C. C. Yang, H. Yang, L. Jiang, and M. Zhang, "Social media mining for drug safety signal detection," in *Proceedings of the 2012 international workshop on smart health and wellbeing*. ACM, 2012, pp. 33–40.

[13] N. Green and S. D. Larasati, "Votter corpus: A corpus of social polling language." in *LREC*, 2014, pp. 3693–3697.

[14] M. Imran, S. M. Elbassuoni, C. Castillo, F. Diaz, and P. Meier, "Extracting information nuggets from disaster-related messages in social media," *Proc. of ISCRAM, Baden-Baden, Germany*, 2013.

[15] H. Shekhar and S. Setty, "Disaster analysis through tweets," in *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on*. IEEE, 2015, pp. 1719–1723.

[16] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, "Processing social media messages in mass emergency: a survey," *ACM Computing Surveys (CSUR)*, vol. 47, no. 4, p. 67, 2015.

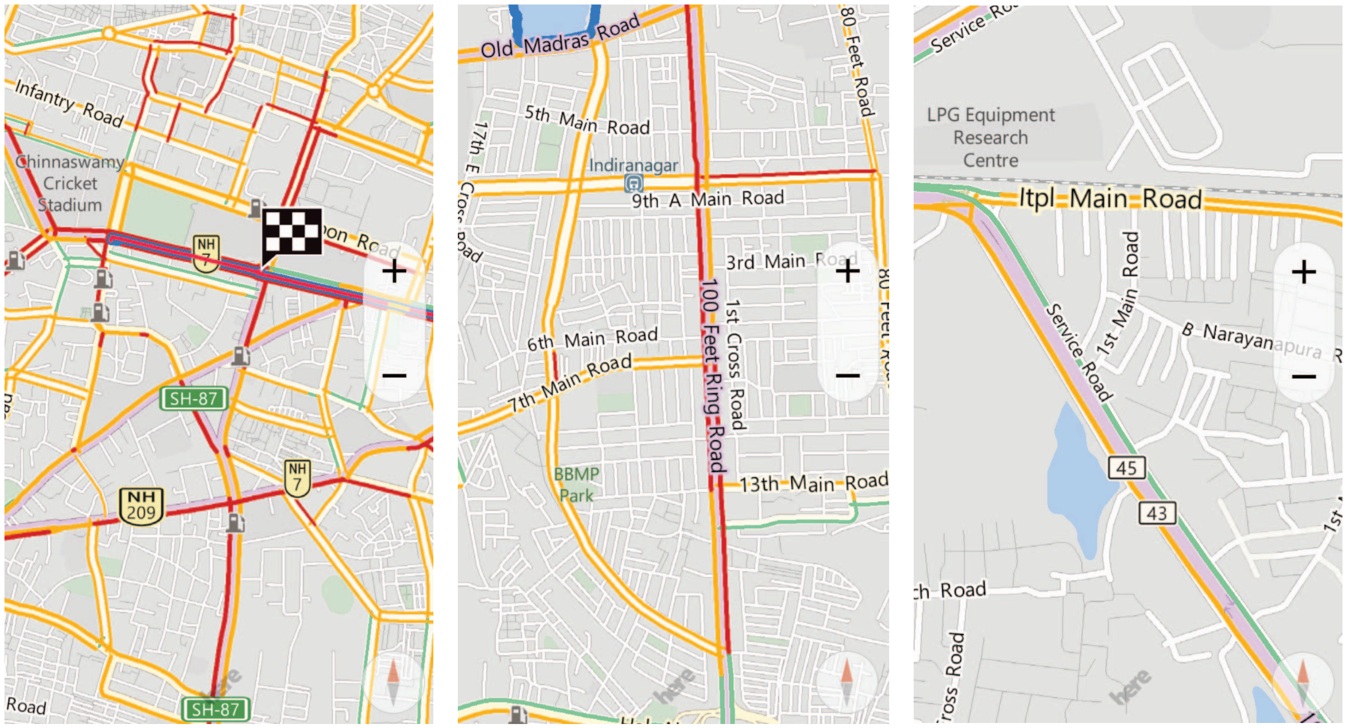
[17] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "Microblogging during two natural hazards events: what twitter may contribute to situational awareness," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2010, pp. 1079–1088.

[18] C. Yang and P. Srinivasan, "Life satisfaction and the pursuit of happiness on twitter," *PloS one*, vol. 11, no. 3, p. e0150881, 2016.

[19] M. Ben-Akiva, M. Bierlaire, H. Koutsopoulos, and R. Mishalani, "Dynamit: a simulation-based system for traffic prediction," in *DACCORS short term forecasting workshop, The Netherlands*. Citeseer, 1998.

[20] W. Min and L. Wynter, "Real-time road traffic prediction with spatio-temporal correlations," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 606–616, 2011.

[21] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Short-term traffic forecasting: Where we are and where were going," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 3–19, 2014.



(a) **Location:** MG Road
Average Sentiment: -6
Event Keywords: cinema, metro

(b) **Location:** Indiranagar
Average Sentiment: -7
Event Keywords: traffic, weekend

(c) **Location:** ITPL
Average Sentiment: -2
Event Keywords: mall, shopping

Fig. 3: Comparison of results with GPS tools. *(a), (b)* are demonstrating severe traffic congestion and *(c)* is demonstrating very moderate congestion in selected localities in Bangalore. Both system's results are consistent. Probable causes of congestion can be speculated through event keywords. *Date: Aug 6th, 2016. Time: 14:30 IST*

[22] A. I. J. Tostes, F. de LP Duarte-Figueiredo, R. Assunção, J. Salles, and A. A. Loureiro, "From data to knowledge: city-wide traffic flows analysis and prediction using bing maps," in *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, 2013.

[23] M. Ni, Q. He, and J. Gao, "Using social media to predict traffic flow under special event conditions," in *The 93rd Annual Meeting of Transportation Research Board*, 2014.

[24] J. He, W. Shen, P. Divakaruni, L. Wynter, and R. Lawrence, "Improving traffic prediction with tweet semantics," in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, 2013.

[25] M. Supriyaningsih, A. Septiani, R. F. Hakim *et al.*, "Mining the traffic conditions via twitter based on rough set theory," in *International Seminar on Innovation in Mathematics and Mathematics Education*. Departement of Mathematics Education Faculty of Mathematics and Natural Science Yogyakarta State University, 2014.

[26] C. P. Khatri, "Real-time road traffic information detection through social media," 2015.

[27] F. Lécué, S. Tallevi-Diotallevi, J. Hayes, R. Tucker, V. Bicer, M. L. Sbodio, and P. Tommasi, "Star-city: semantic traffic analytics and reasoning for city," in *Proceedings of the 19th international conference on Intelligent User Interfaces*. ACM, 2014, pp. 179–188.

[28] F. Lécué, S. Tallevi-Diotallevi, J. Hayes, R. Tucker, V. Bicer, M. Sbodio, and P. Tommasi, "Smart traffic analytics in the semantic web with star-city: Scenarios, system and lessons learned in dublin city," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 27, pp. 26–33, 2014.

[29] S. Grosenick, "Real-time traffic prediction improvement through semantic mining of social networks," Ph.D. dissertation, University of Washington, 2012.

[30] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentistrength, 2011. available online, http://sentistrength.wlv.ac.uk."

[31] M. Thelwall, "Heart and soul: Sentiment strength detection in the social web with sentistrength," *Proceedings of the CyberEmotions*, pp. 1–14, 2013.