Web and Social Media Analytics towards Enhancing Urban Transportations: A Case for Bangalore

Manjira Sinha*
Conduent Labs India
Bangalore, India
manjira87@gmail.com

Preethy Varma
Evestnet Yodlee
Bangalore, India
preethyvarma@gmail.com

Tridib Mukherjee
Conduent Labs India
Bangalore, India
tridibm@gmail.com

ABSTRACT

Cities today are typically plagued by multiple issues such as âĂŞ traffic jams, garbage, transit overload, public safety, drainage etc. Citizens today tend to discuss these issues in public forums, social media, web blogs, in a widespread manner. Given that issues related to public transportation are most actively reported across web-based sources, we present a holistic framework for collection, categorization, aggregation and visualization of urban public transportation issues. The primary challenges in deriving useful insights from web-based sources, stem from:(a) the number of reports; (b) incomplete or implicit spatio-temporal context; and the (c) unstructured nature of text in these reports. This paper provides the text categorization techniques that can be adopted to address specifically these challenges. The work initiates with the formal complaint data from the largest public transportation agency in Bangalore, complemented by complaint reports from web-based and social media sources. An easy to navigate and well-organized dashboard is developed for efficient visualization.. The dashboard is currently being piloted with the largest transportation agency in Bangalore.

Keywords

Text processing, social media analysis, urban informatics

1. INTRODUCTION

Recent years have seen a surge in the web portals and social media discussion forums as well as individual level grievances pertaining to urban transport systems. People are communicating across a broad array of platforms, including dedicated Facebook pages, Twitter handles and hashtags to promote local issues, especially pertaining to transportation. Cities around the world are also using strategies to engage with the residents through social media. While such initiatives can enable residents to provide their opinions and feedback, there is no way for the transport agencies to get meaningful insights (e.g., what are the most talked about problems in which area or routes?) automatically without manually looking into the social media data and complaints.

Typically, transport agencies pre-dominantly rely on complaints center (e.g., call center) for people to provide explicit feedback and complaints manually. Such manual operation is feasible since the number of complaints is less than five per day. On the other hand, the amount of discussions in the web and social media indicates a huge surge in the number of complaints (in the range of hundreds of posts per day regarding transportation problems). Generating meaningful insights from the vast noisy social media data thus become imperative for transportation agencies to understand the pulse of the residents, and subsequently, improve the operations based on the overall feedback. Furthermore, certain types of issues (e.g., no bus service between certain areas in the city) can only exist in the social media. All these observations exacerbates the requirement for automatic analysis of the large-scale web data (in addition to the complaints received the agencies through call center channels) to generate meaningful insights. However, generating such insights is non-trivial and needs to address the following challenges: (i) Categorization of data, (ii) Geo-tagging data, (iii) Route identification, (iv) Report aggregation, and (v) Visualization. To address the aforementioned challenges, we propose an NLP based system to analyze social media data to generate meaningful and actionable insights for transport agencies in urban environment. The proposed system has been tested with the largest transportation agency in Bangalore and the outcome of the system has been verified by the agency.

The rest of the paper is organized as follows. Section 2 presents the proposed system, classification, location tagging, route identification, report aggregation, and visualization module. User interface evaluation details are described in Section 3, section 4 presents the related work, and finally Section 5 concludes the paper.

2. SYSTEM ARCHITECTURE

This section presents the different modules of our system (refer to figure 1) including the description of the input data. The overall, system has six components: preprocessing module, which takes data from the input databases and performs standardization; the next is categorization module that assigns the new and unlabeled posts to appropriate problem categories; the location identification module extracts the place information from the text data; the route identification component detects geographical segment where the problem

*the research has been performed during all authors' stay at Xerox Research Center India.

NDA'17 May 19, 2017, Chicago, IL, USA

DOI: 10.1145/3068943.3068950

is located; the fifth one is aggregation module that clusters the data based on spatio-temporal information; finally, the last component is a visualization dashboard that abstracts on the background complexity to the end user and offer an efficient and easy to use GUI.

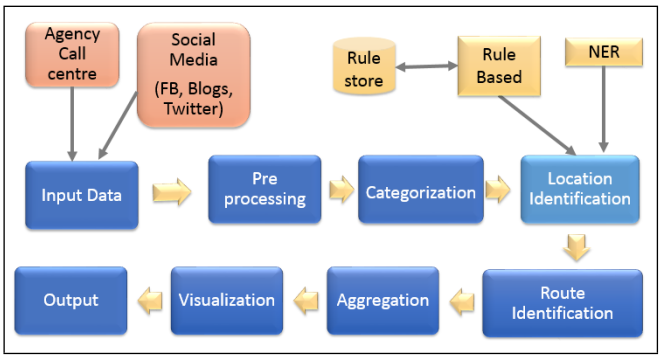

Figure 1: System Architecture

2.1 Nature of the input data

Our primary data source is commuters grievances reported in the largest public transportation agency in Bangalore (referred as the agency from here on). However, to complement the insufficiency of the data, we augment it with publicly available data extracted from various web and social media platforms. The present section describes in detail, the statistics, distribution and nature of the input data.

2.1.1 Data from the Agency

We have obtained 824 user reported problems from the agency's complaints database, collected over a period of 8 months. The reports are categorized across 15 different categories (reduced from 17 after a round of careful examination) such as route deviation, women issues, accidents etc. Refer to figure 2 below.

Category	Example
Crew_Misbehaviour	*"Ms.smitha sridar commuter complaint's that she boarded from banaskari to innovative multiplex when the conductor asked for her daughter to ticket the commuter said that he is 10 years child but the conductor he dosen't look like this and conductor also aggrieve with the commuter"*
Vehicle_Related_Issues	*"500CA KA-57 F-1076 Mr.Vijaykumar Compliant that he traveled from Banashakeri to Whitefelied but the said bus A.C Is Off at 08:10am Pls take action."*

Figure 2: Sample data

2.1.2 Web and Social Media data extraction

As can be observed, the data is too few to build an efficient model. Reporting a complaint in the agency portal requires some formal information such as the Bus ticket number or the bus number. These requirements may attribute to the small number of complaints documented. As a potential remedy, we turn to the web and social media, where the tone is informal and the netizens like to articulate their experiences to the fellow users more often in social media platforms. We crawled user posts from different sources such as public Facebook pages on Bangalore traffic/transport problems, twitter handles on similar issues, and portals like GrahakSeva[1] and IChangeMyCity[2]. To identify posts those are relevant to our context i.e., public transport system in Bangalore, we filter posts containing keywords such as: bus, conductor, Ticket, in addition to the name of the agency. This results in accumulation 2730 posts from the web.

We run a small-scale annotation exercise on the social media posts. Two annotators were consulted to categorize 300 new posts from social media in to the existing 15 categories. However, from the web data, the annotators discover two new categories of problems (refer to figure 3) related to BMTC service. Finally, we have 17 categories for the

Category	Example
Bus_Shelter	*"There are no Bus Shelters on Outer ring Road between Bellandur and Doddanikunte (Total Mall). Makes life its difficult on rainy days"*
No_Bus_Service	*"There is no bus facility in Andrahalli main road i.e near Shiva temple and Anupama High school Areas people are depending upon auto to reach Peenya 2nd stage or Andrahalli bus stop for buses so please provide new bus route at least frequency schedules of 45 minutes"*

Figure 3: New Categories

cumulative over 3000 annotated posts after combining the complaints from the agency database and relevant posts extracted from the social media. Figure 4 presents the final distribution of total 1125 posts from across different classes. Augmenting the complaint data with and web-social media does not only help us in building computational models but also presents to the authority a reflection of the actual state of the transportation problems in the city. Unless otherwise mentioned, henceforth, the word ground truth will indicate this annotated dataset.

2.2 Pre-processing Module

The dataset of short texts typically require some specific pre-processing steps with customized order of execution, in order to receive better quality keywords. Figure 4 presents the steps in this process. As the first step, all proper nouns are removed from the system including a large number of names of people and place names. This makes sure that the names of people or places are not inferred as important keywords. There are many other situations where the names of people/ places are important keywords, for example, if the scenario is to categorize the person or political party from an election campaign, then the proper nouns could be the most important keywords. However, in our scenario, the

[1]http://www.grahakseva.com/
[2]http://www.ichangemycity.com/

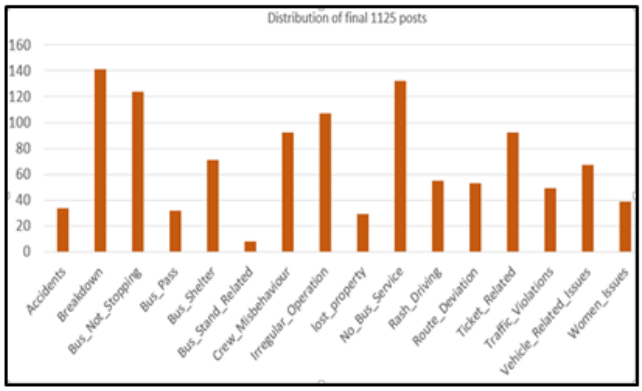

Figure 4: Data distribution

proper nouns carry very less weightage. Tokenization step follows the step of removing extra characters from the text. Extra characters include the URL syntaxes, special characters and numbers. The tokens are passed through a spell correction system, which has two phases. In the first phase, a tweet normalizer tries to normalize the typical representations of words people use on social media. For example, it can convert "plssss"" to "please", "b'tween" to "between". This module uses a dictionary of commonly used social media word representations. The second phase of the spell correction module corrects words by checking the edit distances of words from a set of words kept in a dictionary. After spell correction, the corrected words are passed on to a stop-word filter module that removes the common English stop words. In the next stage, words are stemmed and then domain specific stop words are removed in the next step. The intuitive reason to do a phased stop word removal is that after stemming, different forms of words will converge to a same root form. For example, words like "complaints" and "complain" will have the same root form after stemming and they could be easily removed in the second level of stop word filtering. The domain specific words in this case include, complain, stop, bus, resolve etc. Altogether, there are around 1100 stop words. Once this phase is over, a length based filtering is performed on the tokens. Very short words and very long words are removed. This is because of the fact that such words will be generally very less in occurrence and might not be important. Words shorter than 3 characters and longer than 15 characters are removed. The final step of the pipeline aims at generating n grams, which include the generation of bigrams and trigrams. Finally, the set of unigrams, bigrams and trigrams are presented to the categorizer.

2.3 Categorization Module

Until this point, the unlabeled input data are mere text streams, each as individual entity, without any association. In the categorization module, they are grouped into relevant problem categories by a classifier model developed from the annotated data. We train a support vector machine (SVM) classification model on the 1125-labeled posts. The tf-idf representations of the posts serve as the inputs. The SVM model based on unigram and bi-gram features turns out to be the best performer with cross validation accuracy of 78%. The precision-recall matrix for the data is given below in figure 5. The posts, whose confidence scores are below the threshold are grouped into one separate category others. To evaluate the accuracy of the categorization module, 100 unlabeled data instances are tagged by three annotators. The inter-annotator agreement is measured at 0.75. Against each post, the label assigned by two or more annotators is taken as the problem category. We compare the output of the SVM model with the annotation results and the accuracy of the classifier is found to be 63%. We are in the process of conducting crowd-sourced annotations of a significantly large subset of the public comments to evaluate the correctness of our model and build a sound ground truth.

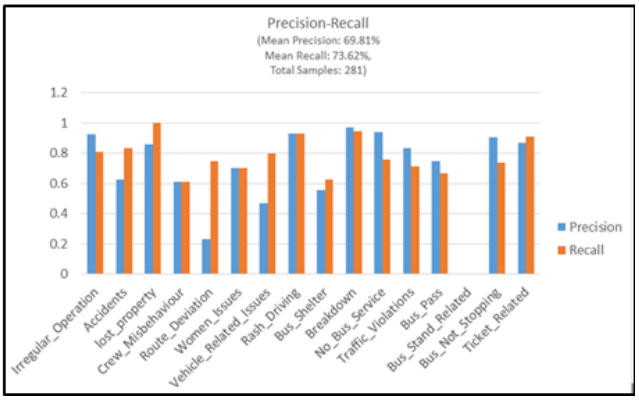

Figure 5: Precision-Recall plot for SVM classifier

2.4 Location Identification Module

Along with classification of the posts in different classes of problems faced by the citizens, we have also grouped the reports according to the problem locations. However, this task of location identification is not trivial. Standard entity identification methods such as named entity recognizer (NER) does not perform well in identifying the place names for short texts [8], due to reasons such as multiple forms of writing place names, spelling mistakes and, short texts like tweets do not follow textual rules of capital casing for place names. For example, a social media post like "in kundalahalli, water has not reached our houses till today" contains the place name "kundalahalli", which is a small region in the city of Bangalore, India. This entity as a possible candidate for place name cannot be recognized by any standard NER. Moreover, maintaining a directory of all possible place names in a particular city is also not a viable option. In this scenario, a rule based system can prove to be very effective and yet simple to develop and maintain. Our rule-based system identify words or phrases, which indicate locations from short texts. The set of rules are generic, therefore, they are not dependent on a dictionary and hence can be applied to a dataset from any location. However, the technique presently is language dependent. Locations are generally represented by the name of a place followed by a location indicator keyword. For example, "AECS layout" contains the location indicator layout. Rules are developed from the textual clues in the posts, through multiple iterations. A set of location indicator words are formed that contains words such as layout. The set is further classified into "Location Indicator Nouns", "Location Indicator Prepositions" and "Location Indicator phrases". Rules heavily depend on the presence and posi-

tioning of the location indicators to identify the locations. With some few exceptions, it is typically observed that one or more indicator words are present in the short text around a location term. Location Indicator Nouns (LIN) are those nouns, which have a proper noun prefix associated with it. For example, "street" or "town" are location indicator nouns. If they are preceded by a proper noun, then it indicates a location or a landmark. In addition, there are location indicator phrases like "residing at", "living near" etc., which indicate that they are followed by some proper noun term, which indicates a place name. The presence of Location Indicator Prepositions (LIP) like "to" and "from" are leveraged to identify location names. An example rule is:

• **"from"** AND **"proper noun/noun"** AND **"towards/to"** IMPLIES **Location = proper noun/noun**.

Refer to figure 6 for sample outputs using our module. Some-

Posts	Locations Identified by NER Toolkits[7]	Locations identified by the rule based system
"Slow moving traffic at Central Silk Board Jn Hosur road towards city due to a bus breakdown"	None	Central Silk Board , Hosur road
"Congested traffic at Jayachamaraja Wodeyar Rd bound from Townhall Circle towards Minerva Circle at PM mirchilavanya suvarnanewstv"	None	Jayachamaraja Wodeyar Rd, Townhall Circle, Minerva Circle

Figure 6: location names identified by proposed module

times it may happen that the location name mentioned in a post refers to a certain landmark (example, name of a shopping mall complex), rather than an actual area name. To standardize the location information, we feed the extracted location names to Google places API [3] and use the first level sub-locality information as a marker for the area containing the location. Each new sub-locality identified is added to a list along with its geographic coordinates. If no sub-locality information is not returned by Google, we look up the latitude-longitude coordinate information corresponding to the specific place name and compare with the sub-locality information in existing list. If any list entry is located within a 5 km radius of the location, we associate the current location with that sub-locality, if no such entry is found the location name is stored as a new sub-locality and flagged for future checking. As the database grows with time, if new sub-locality appears within the distance threshold of 5 km of a flagged location, then the location is merged with the new locality area and the flag is removed. In this way, we dynamically update our location information.

2.5 Route Identification Module

Reports about public transportation system often do not talk only about events occurring at specific locations, but also about a specific transportation routes. When we have any kind of formal information regarding the complaint such as bus ticket numbers, it is easy to map it to a specific route based on the bus running information supplied by the agency. However, such information are hard to find in social media. In order to identify route information from social media posts, we use the input form the location identification module. The location name in a post preceded by terms such as from is considered as source; the location name preceded by terms like to, towards is taken as the destination. Figure 7 is a snapshot from the dashboard featuring the use of the route identification for the posts in the category no bus service. The locations in the right are the destinations and those on the left are the sources, all of these are identified completely based textual contents of the post. Posts containing words like, near and at, on the other hand, points to a fixed location.

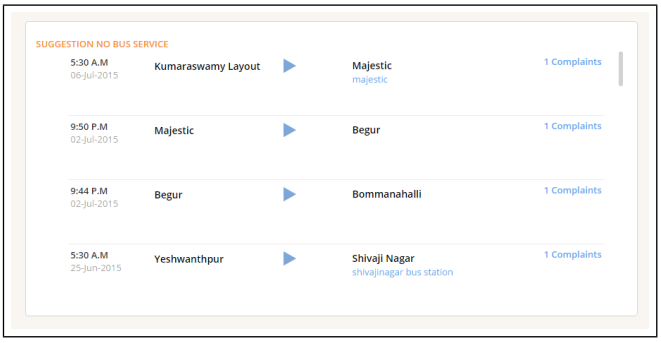

Figure 7: Extracted route information

2.6 Aggregation Module

In web and social media, it is a general trend that multiple posts appear from different users of a particular location to indicate one specific issue. An issue is a problem in a place at a given time, which is reported by multiple people through multiple channels. The issue can be tracked by aggregating all the posts in a specific time window. This also facilitates to track the progress of the problem by comparing with some other time window or locations. Therefore, report aggregation is based on three dimensions viz. location, time and category. As the number of reports in an issue increases, it indicates that the issue is very severe[6]. Two different types of time windows are considered for the temporal aggregation. First, if over an entire day, multiple complaints pertaining to the same category are reported against a location or route, then those posts are grouped as a single issue. For example, if at location A, several bus breakdown complaints are reported on a particular day, then for that day, one of the issues is bus breakdown. However, this issue is not likely to occur soon in the same place. The second type of time-window formulation captures precisely the recurrent pattern of certain issues at a particular location or route. Each day is divided into some time-buckets of size typically 3-4 hours. If corresponding to some location or route, and a fixed time-bucket, posts on the same problem category appear repeatedly over a period time, say 10 days, then aggregation of posts is done. Figure 8 presents a snapshot of the dashboard screen showing the aggregated issues over time.

2.7 Visualization Module

[3]https://developers.google.com/places/

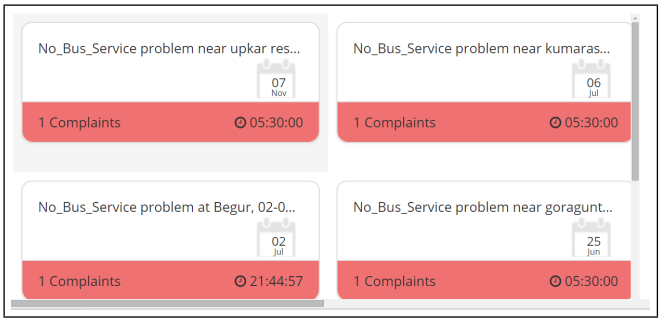

Figure 8: Dashboard display of issues

In addition to processing and management of data, it is also equally important to presents the insights drawn through various techniques in a comprehensible manner. The final component of our system is a dashboard developed to achieve these goals. The GUI is designed in a way that the user will be able to navigate smoothly through the system and features with minimal supervision. The platform provides the following unique offerings to the stakeholder: (i) a single place to view a summary of complaints across various sources; (refer to figure 9 and 10) (ii) ability to slice and dice complaints across locations, over time, and across sources; (iii) get suggestions on new categories (e.g., demand for new routes/services), which the agency does not get through their call center; (iv) ability to prioritize operations according to top regions of complaints and time of year to improve their services as well as leverage new business opportunities (e.g., plying on new routes, increasing number of AC buses during summer, mandating inspections to reduce ticket related issues, and so on). Based on the visualization, the transportation agency can meet the commuter needs more effectively; and in turn can influence increased adoption of public transport in urban environment (thereby reducing congestion and improving overall social sustainability).

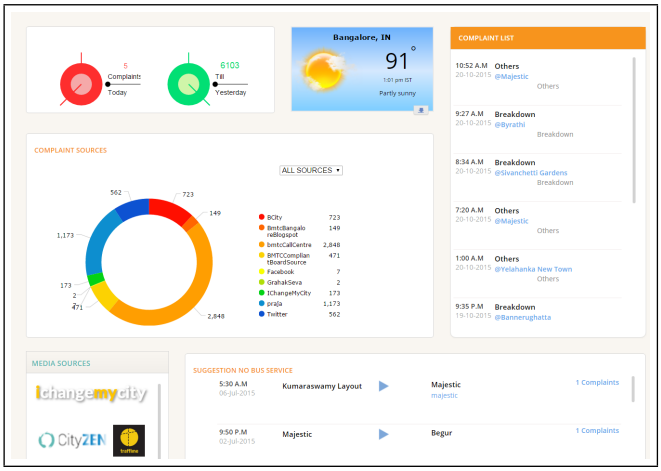

Figure 9: Dashboard Homepage

3. UI EVALUATION

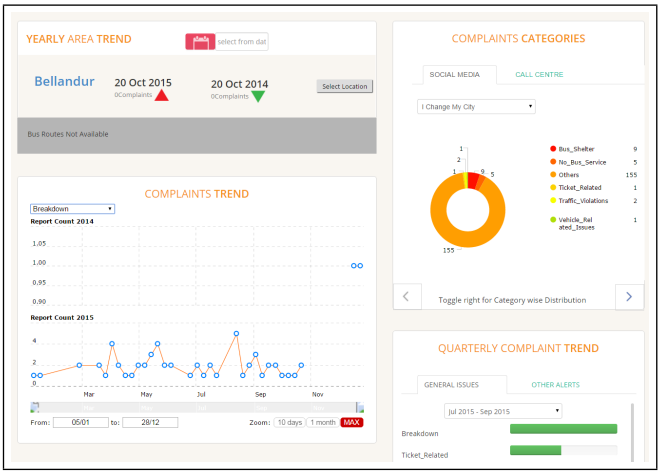

Figure 10: Statistics display in Dashboard

Overall, web service uses elements that make information accessible, readable, and visually appealing for easy daily use. The negative space is abundant and it balances the excess of real time information on the site. Our system attempts to draw insights from the large pool of citizen feedback data available in the social media. In order to substantiate our claim of an easy and efficient visualization of the system, and to identify the scopes for improvement within the navigation, labelling, usability areas, a user evaluation of our dashboard is conducted. The heuristics checkpoints considered in this study are:

• Ease of Primary and Secondary navigation (the logical order)

•Call to action- Will the users know what to click, when or why.

•Texts and content following platform conventions: ensuring that words and buttons mean the same no matter where the user is located at that moment.

The study consists of six (3 male, 3 female) participants, three regular users of the system, two irregular users and one regular monitor of the system.

3.1 Tasks set

The participants are asked to perform the following tasks:

• Navigating into a UI using a screen-shot, (Screen-shot provided was of the issues page âĂŞ action dialog box)

• Find information: Most problematic area in the quarter 1, 2015.

• Find the information: In the category Bus Breakdown, which source provides maximum number of reports?

• Email the report of the issue; “Bus staff rude dated 13.06.2015, 10:08 AM” to a particular e-mail id.

• Find the information: How many complaints in Whitefield, on the date14.07.2015

• Inference Task: From the graphs, infer any trends

As the completion of the above-mentioned tasks requires a user to navigate within a particular page such as Issue page, and from one page to another, we list our observations grouped against three main web pages of the dashboard.

3.2 Observations

• **Homepage:**

Positives: The overall information flow and design are both

simple, effective and direct. The page is clean and focused, containing both textual information and graphs in real time and stored data. High-level information architecture is readily accessible in the form of buttons/ navigations and actions. P1, P2 and P6 accomplished tasks on homepage in lesser times than that taken by the other participants.
Concerns: The section called complaint list on the home page is a real time running data list that scrolls and auto updates; accessing the required information from this list is challenging, as the column does not contain a search bar. The dialog box for selection of an article from a list is noted to be not closing automatically with selection; which could confuse the user as the selected article appears as selected behind the pop up. P4 took the maximum time (4.5 minutes) to accomplish task(s) 2 on homepage, while the average is 3 minutes. Feedback from P1, P2 and P6 state that the negative spacing is ample and soothing and it causes no confusion. Navigation was easy expect for automatic close of dialog box after selection, which is a backend bug. P4 found it difficult to read multiple kinds of visual representation on the same page.
• **Statistics Page:**
Positives: Presentation of all information pages is consistent; graphs (real time data is presented on the graphs) are placed in specific areas on the page and hence, it is accessible and legible. Data is also represented in the form of Bar graphs as quickly inferable quarterly trends that is efficiently inferable. All users went through tasks (3), (5) and (6) on statistics page in less than 2 minutes.
Concerns: No major concerns are noted. However, the excessive amount of various kinds of representation of information may be an information overload for the user. Both the home page and statistics page contain almost equal amount of graphs/ visual information Users P1, P3 and P5 found the tasks (3) (5) (6) harder to accomplish. They were seen to be asking help from the observers. Navigation of task (5) was tougher to accomplish than navigation of task (3) and (6). Feedback state that placement of information is legible. However, the presence of various kinds of bar graphs and charts and textual information on the same page is disrupting data inference.
• **Issues Page:**
Positives: The description of the issue in text alongside the category of complaint and location using a map are the highlight of this page. The information architecture and flow is well defined, with appropriate support information is textual/ visual form. Action buttons to email/ forward are well placed on the footer area of the UI; quickly visible, single click functions. Navigation from homepage to issues and reports dialog is smooth. All users accomplished the task (4) successfully in less than 1 minute. Feedback states that the action buttons on issues page were accessible, legible and had larger click area.
Concerns: The only concern put forth by P6 is the lack of prompting an action.

4. RELATED WORKS

At present times, various service-providing agencies are using mobile crowdsourcing, crowd sensing, and human participatory sensing systems to explore the possibility of collecting implicit and explicit feedback from the residents on urban issues [6, 7, 1], so that meaningful insights can be extracted. However, such systems, due to their formal requirements often results in lower and slower penetration of within the users. It is thus important to complement these sources with the social media feeds and other online sources to generate rich insights on city-related issues. A number of studies have come in recent times that leverage the power of social media to detect public events and infer public sentiments corresponding to various issues. In [2], the authors have tapped into the real time messages shared on Twitter to identify seismic activities like earthquakes, by applying burst detection techniques on twitter data. Language independent techniques for event detection have been developed based on user behavior modelling in social media [4]. The authors have noted that communications among users drop while some important external event is occurring. Social media data have also been used for traffic prediction in long term [5]. LASTA [9] is a good example of using social media tools in production systems. It is a topic mining system being used as a production level application to mine topics from multiple online sources and assigning them to large number of topics. Although analysis of social media to infer topics of interest, sentiments, etc. is well studied, these studies do not correlate information from other heterogeneous channels such as mobile app, public blogs, web pages, emails etc. Typically, city agencies have dedicated departments to handle complaints where complaints are received via phone calls, emails, mobile apps, and online portals. The complaints are then handled manually in the current state of the art. However, aggregation of multiple complaints across multiple channels of information for each issue, and then getting actionable insights, can lead to effective decision-making and city management at a very large scale. Spatio-temporal clustering of social media data is a topic in recent focus [3]. However, generating clusters (of reports to identify distinctive issues) online while being aware of specific landmarks as well as existing issues in the city is unexplored. This paper attempts to fill in this gap by correlating and aggregating information from heterogeneous channels such as mobile phones, social media, and other online sources based on spatial, temporal, and topical relevance. Such association to landmarks is essential to generate actionable insights for city agencies, where problems make more sense with reference to a landmark, rather than lat-long coordinates.

5. CONCLUSION AND FUTURE DIRECTION

This work is a crucial first step towards helping different administrative bodies in better management of urban infrastructure related issues through insights on public perception. Apart from the administrative bodies, common citizens will also be benefited from this kind of work as they can visualize and understand the city as a whole. Our long-term objective is to establish a trusted channel of communication through various social media analysis techniques, between the citizens and city administration, in order to realize the dream of smarter, empowered and shared cities.

6. REFERENCES

[1] H. Abdelhaq, C. Sengstock, and M. Gertz. Eventweet: Online localized event detection from twitter. *Proceedings of the VLDB Endowment*, 6(12):1326–1329, 2013.

[2] M. Avvenuti, S. Cresci, A. Marchetti, C. Meletti, and M. Tesconi. Ears (earthquake alert and report system):

a real time decision support system for earthquake crisis management. In *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1749–1758. ACM, 2014.

[3] M. Budde, J. De Melo Borges, S. Tomov, T. Riedel, and M. Beigl. Leveraging spatio-temporal clustering for participatory urban infrastructure monitoring. In *Proceedings of the First International Conference on IoT in Urban Space*, pages 32–37. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2014.

[4] F. Chierichetti, J. M. Kleinberg, R. Kumar, M. Mahdian, and S. Pandey. Event detection via communication pattern analysis. In *ICWSM*, 2014.

[5] J. He, W. Shen, P. Divakaruni, L. Wynter, and R. Lawrence. Improving traffic prediction with tweet semantics. In *IJCAI*, 2013.

[6] T. Mukherjee, D. Chander, S. Eswaran, M. Singh, P. Varma, A. Chugh, and K. Dasgupta. Janayuja: A people-centric platform to generate reliable and actionable insights for civic agencies. In *Proceedings of the 2015 Annual Symposium on Computing for Development*, DEV '15, pages 137–145, New York, NY, USA, 2015. ACM.

[7] A. Nurwidyantoro and E. Winarko. Event detection in social media: A survey. In *Proceedings of the International Conference on ICT for Smart Society (ICISS), IEEE*, pages 1–5, 2013.

[8] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[9] N. Spasojevic, J. Yan, A. Rao, and P. Bhattacharyya. Lasta: Large scale topic assignment on multiple social networks. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1809–1818. ACM, 2014.