Report No.	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle

User Sentiment Analysis with Louisiana Social Media Data for Effective Crash Countermeasures

5. Report Date
June 2015

6. Performing Organization Code
LTRC Project Number:
SIO Number:

7. Author(s)
Subasish Das
Xiaoduan Sun, Ph.D, P.E.

8. Performing Organization Report No.
University of Louisiana at Lafayette

9. Performing Organization Name and Address

Department of Civil and Environmental Engineering
University of Louisiana at Lafayette
Lafayette, LA 70504

10. Work Unit No.

11. Contract or Grant No.

12. Sponsoring Agency Name and Address
Louisiana Department of Transportation and Development
P.O. Box 94245
Baton Rouge, LA 70804-9245

13. Type of Report and Period Covered
Final Report
June 2015

14. Sponsoring Agency Code

15. Supplementary Notes
Conducted in Cooperation with the U.S. Department of Transportation, Federal Highway Administration

The microblogging platform Twitter, with nearly 600 million users and over 250 million tweets per day, has become one of the most popular quick and efficient information sharing platforms in recent years. The users can share their instant thoughts or information on a wide range of topics or interests through short messages known as 'tweets'. Meaningful extraction of information from the large amount of data on Twitter requires elegant scientific data-extracting tool designs. Corporations, news agencies, transportation authorities, local governments, airlines and various other agencies that share real-time public service information use Twitter. The prospect of using social media to develop links between citizens and governmental authorities is an evolving issue in theory and practice.

This project examines the promising aspects of the governmental transportation authorities in providing information to the people. Based on a dataset of nearly nine thousand official tweets from transportation authorities in two major Louisiana cities, the researchers explored how local authorities attempted to provide information to roadway users through Twitter. Researchers perform semantic analysis on the tidy dataset to extract knowledge. This study demonstrates how text mining retrieves knowledge from official transportation information tweets. It is clear that Twitter usage in governmental information sharing is significant which enables users to share important information under extreme weather or other calamities. The frequent term analysis from both of the Twitter handles is similar. The tweets are mostly related to terms like 'congestion', 'blocked', 'lane/lanes', 'accident' and 'open'. Correct usage of these tweets is usually economically and environmentally beneficial for both transportation agencies and road users. Moreover, we also performed sentiment analyses on a few countermeasures in this research. The results showed mixed sentiments towards these countermeasures. Authorities need to examine negative sentiments more profoundly to improve performance.

17. Keywords
AADT, traffic volume, rural local roadways, machine learning, pattern recognition, support vector machine, support vector regression

18. Distribution Statement
Unrestricted. This document is available through the National Technical Information Service, Springfield, VA 21161.

20. Security Classif. (of this page)	20. Security Classif. (of this page)	21. No. of Pages:	22. Price

Project Review Committee

Each research project will have an advisory committee appointed by the LTRC Director. The Project Review Committee is responsible for assisting the LTRC Administrator or Manager in the development of acceptable research problem statements, requests for proposals, review of research proposals, oversight of approved research projects, and implementation of finding.

LTRC appreciates the dedication of the following Project Review Committee Members in guiding this research study to fruition.

LTRC Administrator

Kirk M. Zeringue, P.E.

Members

User Sentiment Analysis with Louisiana Social Media Data for Effective Crash Countermeasures

by

Subasish Das
Ph.D. Student

Xiaoduan Sun, Ph.D., P.E.
Professor

Civil Engineering Department
University of Louisiana at Lafayette
254 Madison Hall
100 Rex Street
Lafayette, LA 70504

LTRC Project No.
SIO No.

Conducted for

Louisiana Department of Transportation and Development
Louisiana Transportation Research Center

June 2015

ABSTRACT

Twitter, with nearly 600 million users and over 250 million tweets per day, has become one of the most popular microblogging platforms in the recent years. The users can share their instant thoughts or information on a wide range of topics or interests through short messages known as 'tweets'. Meaningful extraction of information from the large amount of data on Twitter is a great challenge. Corporations, news agencies, transportation authorities, local governments, airlines and various other agencies that share real-time public service information use Twitter. The prospect of social media for cooperation between citizens and governmental authorities is an evolving issue in theory and practice. This project examines the promising aspects of the governmental transportation authorities in providing information to the people. Based on a dataset of nearly nine thousand official tweets from transportation authorities in two major Louisiana cities, the researchers explored how local authorities attempted to provide information to roadway users through Twitter. Semantic analysis performed on the neat and tidy dataset helps extract knowledge from the tweets. This study demonstrates how text mining retrieves knowledge from official transportation information tweets. It is clear that Twitter usage in governmental information sharing has a significant impact, which enables users to share important information under extreme weather or other calamities. The frequent term analysis from both of the Twitter handles is similar. The tweets are mostly related to terms like 'congestion', 'blocked', 'lane/lanes', 'accident' and 'open'. Real-time usage of these tweets will be economically and environmentally beneficial for both transportation agencies and road users. Moreover, the researchers examined sentiment analyses on a few countermeasures in this research with the results showing mixed sentiments towards these countermeasures.

ACKNOWLEDGMENTS

The research team wants to thank the LTRC project review committee and UL Lafayette undergraduate student work Leia Kagawa.

IMPLEMENTATION STATEMENT

This research has developed a Twitter analysis and sentiment analysis tool to extract transportation-related knowledge from location-specific tweets. The transportation authorities can conduct research on systems and policy performance by using the codes and tools developed in this research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Improving highway safety is a critical issue facing DOTD because the state's traffic fatality rate (fatalities per 100 million vehicle miles traveled) has been consistently higher than the national average despite the improvements made in the last several years. In 2011, the national average fatality rate was 1.13 while Louisiana had 1.54. To reach "Destination Zero Deaths" set by Louisiana Highway Safety Strategies, it is critical to reduce the number of crashes and crash severities.

Recent years have witnessed a surge of interest in computational methods for affect, ranging from opinion mining, to subjectivity detection, to sentiment and emotion analysis. These methods typically focus on the identification of private states, such as opinions, emotions, sentiments, evaluations, beliefs, and speculations in natural language. While subjectivity classification labels text as either subjective or objective, sentiment classification adds an additional level of granularity by further classifying subjective text as positive, negative or neutral. The goal of opinion mining is to identify emerging societal trends based on views, dispositions, moods, attitudes and expectations of stakeholder groups or the public. This approach definitely helps in the area of policymaking to better anticipate likely impacts of policy measures and better communicate expected benefits and consequences.

Social media resources like Facebook and Twitter generate immense amounts of textual data on various topics. With the tremendous growth of these networks, there has been a growth of data generation every minute on these networking sites. The data extraction on particular interests (like a newly added traffic law) or the countermeasures (like child restraint usage, safety-belt usage, street lighting or red light cameras) can help the policy makers see user sentiment on those items. This study will use the Louisiana-based social media data to investigate the user opinions and sentiments towards an interest group set of countermeasures.

Analyzing large textual data helps corporations and governments understand the public sentiment regarding business, marketing, academic, and policy-making strategies. Usage of Twitter is not only limited to the public. Corporations, news agencies, transportation authorities, airlines and various other agencies that share real-time official information also use it. The prospect of social media usage for cooperation between citizens

and governmental authorities is an evolving issue in theory and practice.

We know that real-time traffic information helps roadway users make smart decisions. Twitter is an immediate and flexible tool to broadcast travel information and communicate through brief public messages. The Louisiana Department of Transportation and Development (DOTD) maintains 16 Twitter accounts to deliver traffic information to the public. The DOTD official Twitter accounts have nearly 53,255 followers and 83, 641 tweets were posted from these accounts as of June 20, 2015. It is interesting to know the users' responses to these tweets and the impact of tweeting on the information dissemination. There are few studies conducted to investigate the effectiveness of social media usage by the governmental traffic authorities. Based on a dataset of nearly fourteen thousand official tweets from two major city traffic authorities in Louisiana, this research explored how local authorities provide information to roadway users through Twitter and their potential developments. This project also conducted sentiment analyses on the DOTD official Twitter handles as well as on some significant countermeasures by conducting a term search in Twitter.

Literature Review

Twitter generates short messages that provide both public sentiment and real-time necessary information. In recent years, many researchers have conducted research on Twitter analysis. Twitter analysis generates many kinds of research: sentiment analysis and opinion mining, algorithmic improvement of sentiment scores, keyword extraction, knowledge discovery, policymaking strategies, political prediction and many other related researches. Even though literature on the use of Twitter mining has expanded in recent years, there is still a lack of studies on the use of Twitter by local governmental authorities.

Automated extraction of important information is an important area of text analytics. Text mining is a component of natural language processing (NLP) and it aims to convert text into intuitive knowledge. The process extracts information from text by applying statistical or machine learning algorithms [1-2]. Information extraction is followed by analysis of the retrieved and transformed structured data to specify clusters and relationships between different perceptions. Many researchers value the importance of the text mining approach as a way of improving the accuracy of data collection. Semantic analysis of the social media

data was widely used to facilitate many applications: user interest modeling [3], sentiment analysis [4], content exploration [5-7], event tracking [8], citizen-government relations [9-11], news retrieving [12], prediction of stock market variations [13], the management of natural disasters [14], the understanding of epidemical diseases [15] and the characterization of electoral processes [16]. The authors compiled a more detailed bibliography (with abstracts of the papers) on social media research in a webpage [17].

The internet contributes a lot in delivering better public services by rendering the association between citizens and governmental authorities to revitalize public service. There is a need for research to identify what extent the usage of social media tools satisfies the public demand and to what extent these tools contribute to improve the system performance. The study on the impact of social media usage of the governmental traffic authorities of Louisiana is therefore called for. As the first step of this research, we address the responsiveness of the governmental traffic authorities to traffic conditions by providing real-time traffic information to the traveling public. Moreover, we conduct Louisiana-specific sentiment analysis on a few countermeasures in practice in the state.

OBJECTIVE

The goal of this project is to develop a better algorithm for extracting people's sentiments on a particular interest group set. This research also aims to examine the promising aspects of the governmental transportation authorities in providing information to the people. Specifically, the objectives of the proposed project are to:

- Explore data analysis on the collected yearly tweets from two major Twitter handles of DOTD.
- Perform text mining on the collected tweets.
- Develop algorithms for better text refining and knowledge distillation from large sets of social media data.
- Add transportation safety related terms in senti-lexicon.
- Conduct sentiment analyses on DOTD tweets and tweets related to specific countermeasures.

SCOPE

The research aims to develop a methodology to understand people's sentiments as well as examine the promising aspects of the governmental transportation authorities in providing information to the people.

METHODOLOGY

Text mining is an applied method that originated from a more generic scientific branch called data mining or knowledge discovery. Knowledge discovery is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [18]. We view Knowledge Discovery in Text (KDT) or text mining as a multi-stage process that comprises all activities from document collection to knowledge extraction. It utilizes approaches like data mining, information retrieval, supervised and unsupervised machine learning and computational semantics. Extraction of useful information from data resources through pattern recognition helps identify contributing factors in associated tasks. Text mining mainly deals with the collections of unstructured textual data rather than from structured databases.

Twitter is a relatively new social media tool for microblogging. The user posts, known as tweets, do not exceed 140 characters without any privacy conditions. Therefore, it not only disseminates information but also reflects opinions in real-time. Some information and unfiltered opinions can be very sensitive in various aspects. Twitter generates a huge amount of textual content daily. We can study textual content by means of text mining, natural language processing, information retrieval, and other methods. It is true that there is an open debate on whether Twitter stratifies the necessary representative sample data of the outside world.

However, a contextualization of social media data with an appropriate mechanism may provide important insights. The keys to successful Twitter mining depend on several factors, such as appropriate algorithms, target specification, and responsiveness of the post. The definitions related to Twitter are described here in brief for familiar interpretation:

Tweet: A short message, post or microblog from an account holder on Twitter. The account holder's identification name is known as a Twitter handle. The text spans a maximum of 140 characters. Tweets include updates about activities, useful information; forward of other's tweets, conversations, etc.

Hashtag: Denoted by a word with preceding '#' symbol (e.g., #NOLA_Traffic). It is generally used before a relevant keyword or phrase (no spaces) in tweets to categorize those tweets and help them show more easily in Twitter Search.

Reply: The Twitter feature reply helps in responding to a tweet. This syntax automatically inserts the originator's user name.

Retweet: Retweet forwards a tweet from users to their followers. It is almost similar to e-mail

forwarding.

Mention: Mention acknowledges a user with the symbolic '@' sign without using the reply feature.

We divide the methodology into four major tasks: 1) Data collection, 2) Data preparation and exploratory data analysis, 3) Text mining, and 4) Sentiment analysis.

Data Collection

In information retrieval approaches, we assume keywords as a representation of compact information in documents. Keyword extraction uses a natural language processing method to identify particular word/term tags combined with supervised or unsupervised machine learning algorithms. Moreover, calculations on co-occurrence of particular phrases and terms would be a point of interest in various researches. For example, high frequency of the term 'congestion' would indicate the nature of the document's particular interest. If the occurrence of 'congestion' with another term 'minimal' were high, it would rather indicate a different nature of the document's interest. In text mining, corpus represents a collection of text documents. A corpus is an abstract concept, and there can exist several implementations in parallel. After developing a corpus, users can easily modify the documents in it: stemming, stop word removal, numbers, particular parts of speech, and redundant words are all examples of this. The flowchart of the Twitter mining approach developed in this study is shown in Figure 1.

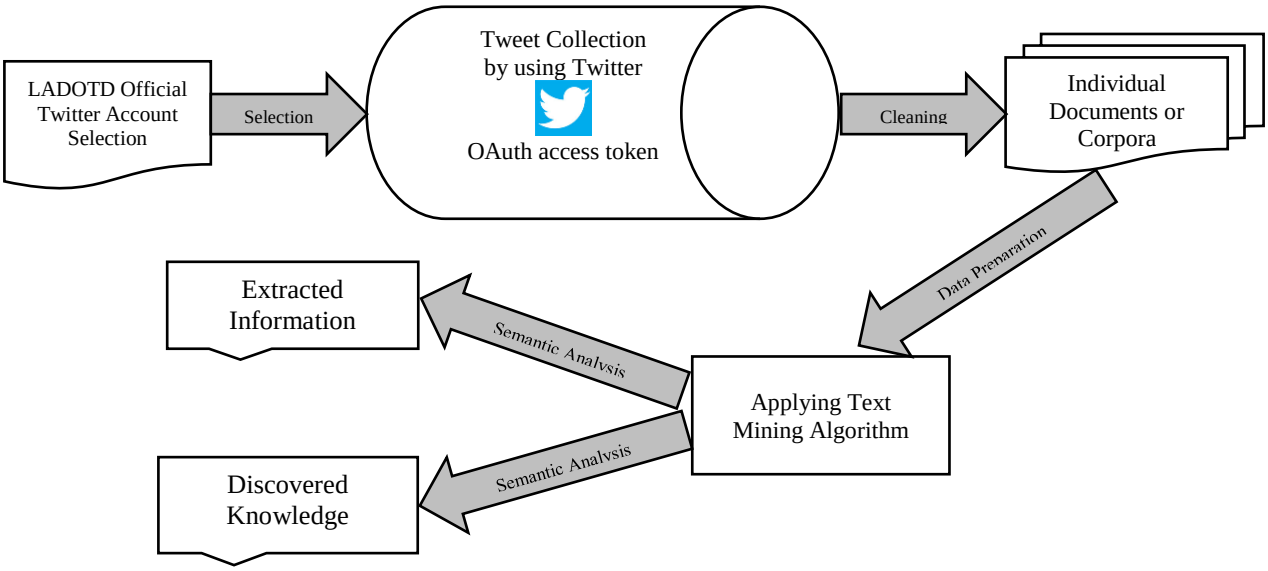


FIGURE 1 Flowchart of text mining

Improved public service constitutes the vital part of the administrative performance of a government office. At present, the traditional concept of public service delivery is changing drastically. Because of the increasing demand for diverse social needs and reliable public services, the governmental authorities have to take cordial steps to enhance their services despite the budgetary constraints. Social media has become a potential tool to make significant contributions in providing better public services.

DOTD maintains sixteen official Twitter accounts. The official Twitter handles for Baton Rouge and New Orleans are the most dominant ones in number of tweets and followers among them [Table 1]. The researchers collected tweets from both of these Twitter handles (BR_Traffic and NOLA_Traffic). Both of these official accounts were created in January of 2009. Twitter currently implements two forms of authentication in the new model, both still leveraging open standard for authorization (OAuth). These two forms are: 1) Application-user authentication that is the most common form of resource authentication in Twitter's OAuth 1.0A implementation to date. 2) Application-only which is a form of authentication where user application makes API requests on its own behalf, without a user context [19]. It is important to note that the one-time tweet extraction limit from a Twitter handle is 3,200.

TABLE 1 Official Twitter Accounts of LADOTD

City	Official Twitter Handle	Tweets	Followers
New Orleans	NOLA_Traffic	34,700	21,300
Baton Rouge	BR_Traffic	26,200	27,900
Geauxwider	GeauxWider	7,469	886
Shreveport	Shreveport_Traf	5,389	2,464
North Shore	NS_Traffic	3,340	1,657
Houma	Houma_Traffic	2,568	1,776
Lafayette	Laf_Traffic	1,862	1,292
Lake Charles	LC_Traffic	1,385	373
Monroe	Monroe_Traffic	432	264
Geauxpass	GeauxPass	179	222
Alexandria	Alex_Traffic	117	178
	Total	83,641	58,312

We used popular data mining ‘R’ packages “twitteR” and “tm” in this study to extract tweets from the user timeline of two official DOTD Twitter handles and semantic analysis respectively [20]. The collected tweets from both of the handles were for an eight-month period. After collecting the tweets, we divided the text of the tweets into four different documents or

corpora per Twitter handle. The division considers the time stamp hour of the tweets as a basis of separation 12AM-6AM, 6AM-12PM, 12PM-6PM, 6PM-12AM. We implemented different Twitter mining approaches on the dataset to get the insight of the textual data.

Data Preparation and Exploratory Data Analysis

The total number of tweets analyzed in this research is nearly fourteen thousand (only 2014 tweets are counted). The official tweets were retweeted by their followers nearly 51,000 times. Figure 2 shows the tweets and retweets generated from these accounts. In terms of number of retweets, the followers of NOLA_Traffic retweeted nearly twice that of the BR_Traffic followers. The researchers built a webpage with a list of all official tweets posted from these two Twitter handles [21]. The peak of the retweets from both handles is visible on January 25, 2014 when transportation authorities closed the interstates due to severe icy conditions. Both of the Twitter handles shared the most recent status by tweeting real-time information. The followers retweeted those tweets to inform their own followers. This event clearly shows the necessity of using Twitter for information dissemination and sharing by the transportation authorities. Benefit-cost analysis from the social media usage for this particular case can be explored as a prospective future research topic.

How each document (based on hour of the day and month of the year) represents the percentage of all tweets is shown in Table 2. We performed the final analysis on the hourly-based tweets. The unique terms used in the tweets are 570 on average. Term-document matrices tend to become large enough for normal-sized data sets so it is required to remove sparse terms (terms occurring only in very few documents). Sparsity of the terms generated from all tweets was nearly 50 percent. By removing the sparse terms, the matrix dramatically reduces without losing significant relations inherent to the matrix. In this study, we removed 25% of sparse elements to make the document more intuitive and noiseless. After removing the sparse terms, the sparsity of the document reduced to 0%.

TABLE 2 Descriptive Statistics

	Baton Rouge	**New Orleans**
Official Account Name	BR_Traffic	NOLA_Traffic
Analyzed Tweets	3304	5605
Hour of the Day		
12AM- 6AM	15.56%	13.95%
6AM-12PM	13.01%	17.47%
12PM-6PM	35.71%	38.93%
6PM- 12AM	35.71%	29.65%
Months		
November	11.29%	12.88%
December	13.14%	11.79%
January	12.89%	12.79%
February	11.99%	12.52%
March	14.62%	11.83%
April	10.87%	12.38%
May	13.11%	12.85%
June	12.11%	12.95%
All Tweets		
Terms	569	571
Documents (based on Hours)	4	4
Non/sparse entries	1163/1113	1192/1092
Sparsity	49%	48%
Maximum Term length	21	21
After Removing Sparse Terms (0.25)		
Terms	109	116
Documents (based on Hours)	4	4
Non/sparse entries	436/0	464/0
Sparsity	0%	0%
Maximum Term length	12	16

Excluding the redundant terms (combined frequency lower than 100, names of the streets, numbers, specific parts of speech, article etc.), in the eight-month period during which the study was conducted, the frequency of the terms is shown in Figure 3. In BR_Traffic the top five highly frequent terms are congestion, lane/lanes, blocked, open, accident/accidents. In NOLA_Traffic the top five highly frequent terms are lane/lanes, congestion, blocked, open, and minimal. Another visual representation of the cleaned Twitter data is shown in Figure 4 based on the generated corpus. The heat map clearly identifies the significance of the terms based on the hour of the tweet posts. The findings from Figure 4 also match with the visual display of Figure 3.

We listed the findings obtained from the tweet content analysis in Table 3. Presence of a particular term will be more intuitive if one knows what the most correlated terms it comes with are. The correlation ratio for the terms associated with three important terms, which are 'congestion', 'blocked', and 'accident', are given in Table 3. We assumed the least correlation factor as 0.97. When 'congestion' is associated with 'minimal', it implies a less congested phase. The term 'minimal' is highly correlated with 'congestion' in the NOLA_Traffic handle while it is less correlated with 'congestion' in the BR_Traffic handle. This particular case implies the traffic congestion condition for both of the cities.

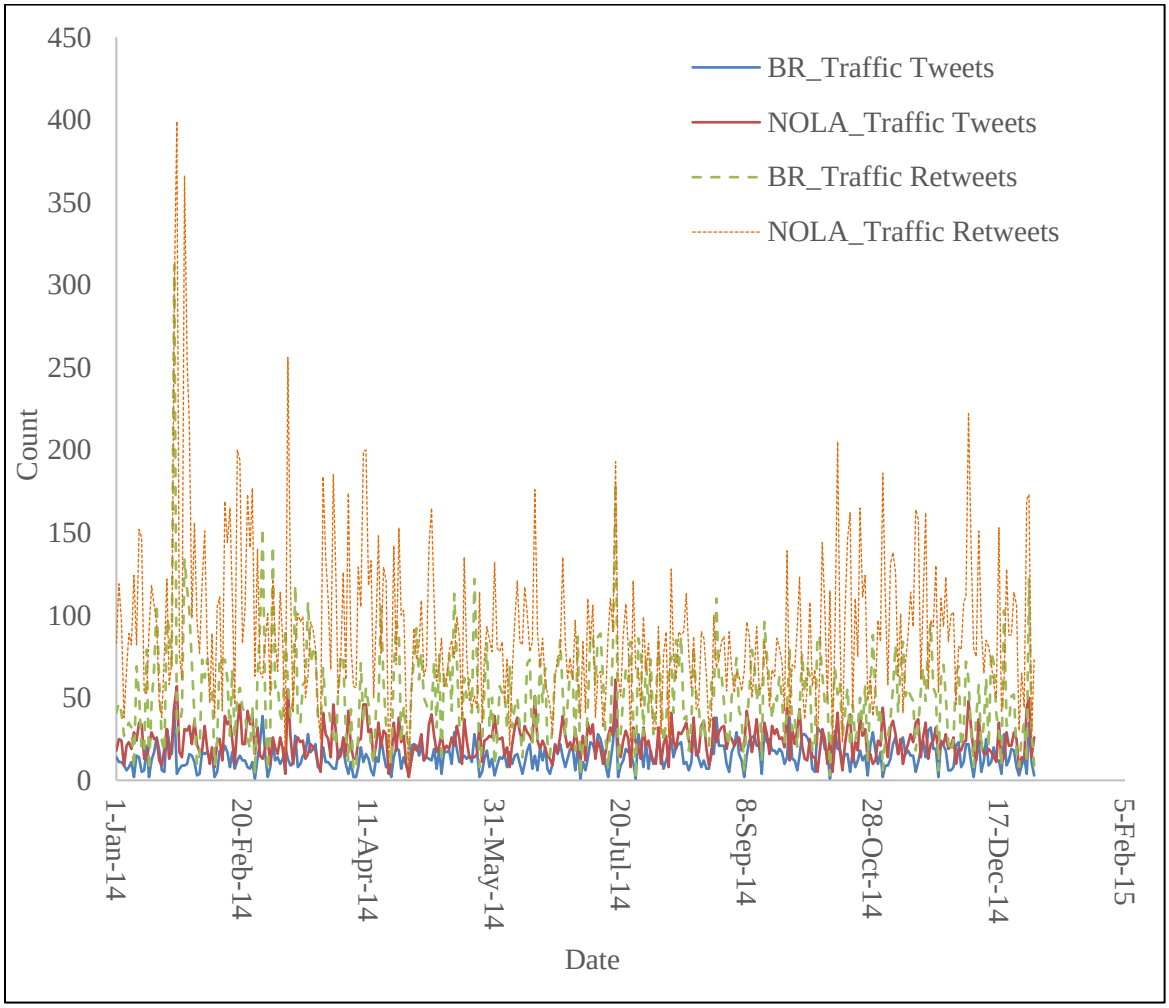


FIGURE 2 Count of tweets and retweets

Text Mining

In the hierarchical clustering tree structure (dendrogram) [Figure 5], the terms are listed along the y-axis. The x-axis measures inter-cluster distance. Ward's hierarchical clustering considers

TABLE 3 Correlations between Terms

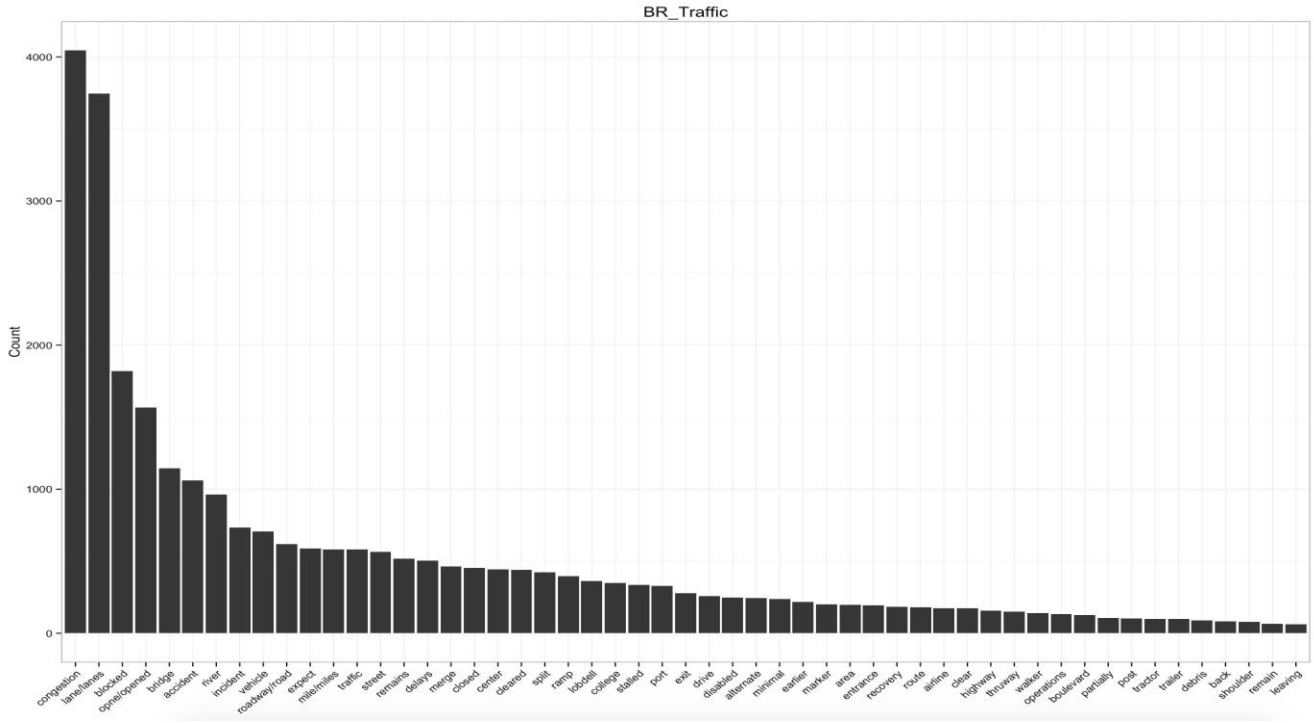
Baton Rouge		New Orleans	
Congestion		**Congestion**	
Lane	1.00	Blocked	1.00
Open	1.00	Connection	1.00
Vehicle	1.00	Lane	1.00
Blocked	0.99	Veterans	1.00
Lanes	0.99	City	0.99
Miles	0.99	Lanes	0.99
Port	0.99	Open	0.99
Post	0.99	Split	0.99
Old	0.98	Vehicle	0.99
Accident	0.97	Center	0.98
Cleared	0.97	Opened	0.98
Hwy	0.97	Truck	0.98
Length	0.97	Accident	0.97
Minimal	**0.63**	Minimal	**0.92**
Blocked		**Blocked**	
Port	1.00	Congestion	1.00
Vehicle	1.00	Connection	1.00
Congestion	0.99	Lane	1.00
Hwy	0.99	Center	0.99
Lane	0.99	City	0.99
Accident	0.98	Disabled	0.99
Merge	0.98	Lanes	0.99
Post	0.98	Open	0.99
Clear	0.97	Split	0.99
Open	0.97	Truck	0.99
Weather	0.97	Vehicle	0.99
		Veterans	0.99
		Opened	0.98
Accident		**Accident**	
Clear	1.00	Remains	1.00
Overpass	1.00	Causeway	0.99
Vehicle	0.99	Shoulder	0.99
Blocked	0.98	Split	0.99
Center	0.98	Vehicle	0.99
Currently	0.98	Leaving	0.98
Old	0.98	Passing	0.98
Congestion	0.97	Congestion	0.97
Hwy	0.97	Pkwy	0.97
		Through	0.97

each term as unique before merging it with the most similar terms and then merging the resulting pairing with the most similar pairing and so on. The resulting dendrogram offers a powerful summary of the aforementioned analysis and confirmation of the association between different term characteristics. For example, congestion and lane/lanes have the same inter-cluster distances. ‘Congestion, lane/lanes’ are associated with another group of terms (accident, bridge, minimal, open, and blocked). The term ‘route, recovery’ has the smallest inter-cluster distance in Figure 5(a) and ‘leaving, passing’ have the smallest inter-cluster distance in Figure 5(b). The closely grouped terms have some indication of association of the tweets.

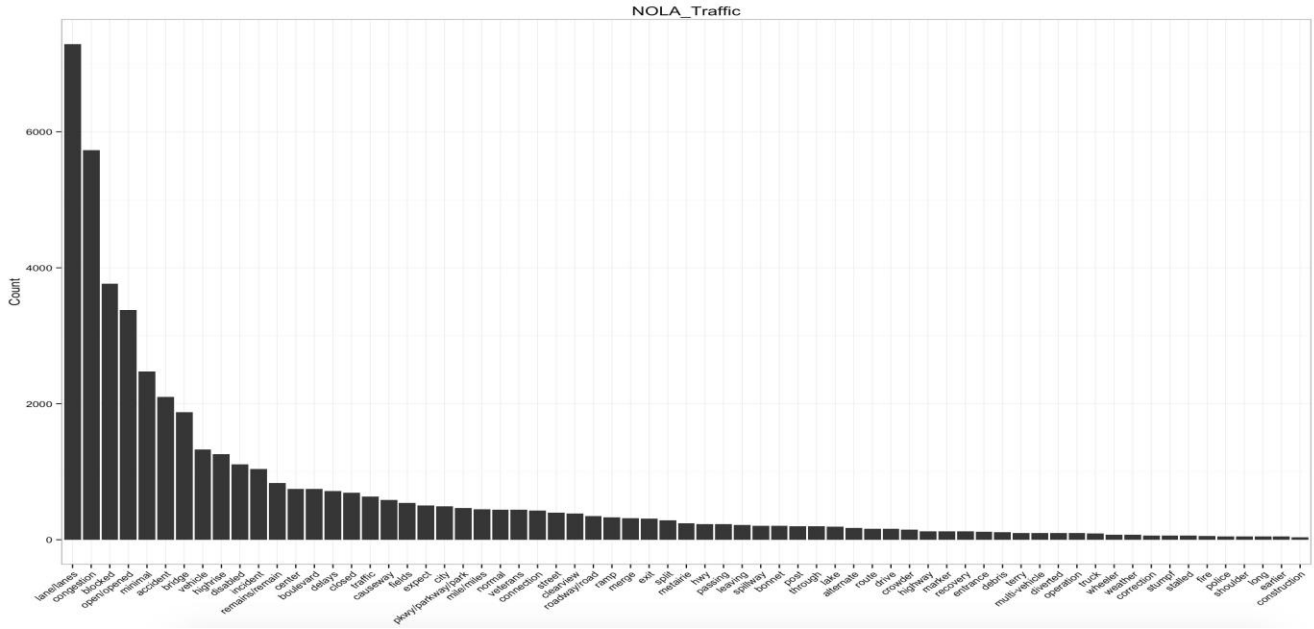
The findings from the analysis are:

- Social media helps in improving public service during bad weather. Real-time information on roadway blockage and other travel problems would be economically beneficial to traveling public.
- Text mining shows that most of the tweets were associated with terms like ‘congestion’, ‘blocked’, ‘accidents’, ‘lane/lanes’, and ‘open’. Real-time utilization of these terms would lessen highway mileage and would be environmentally and economically beneficial.
- The hour-based heat map implies the percentage of presence of the terms used in different documents. The findings are similar with the research analyzing peak-hour traffics.
- The tweets related to Baton Rouge Traffic informs more about congestion, while congestion is highly associated with the term ‘minimal’ in the tweets for New Orleans traffic.

The limitation of this study is the usage of limited data (only 2014 Twitter data). The complete analysis of the tweets since January 2009 would be a good analytic approach to extract knowledge from the data.

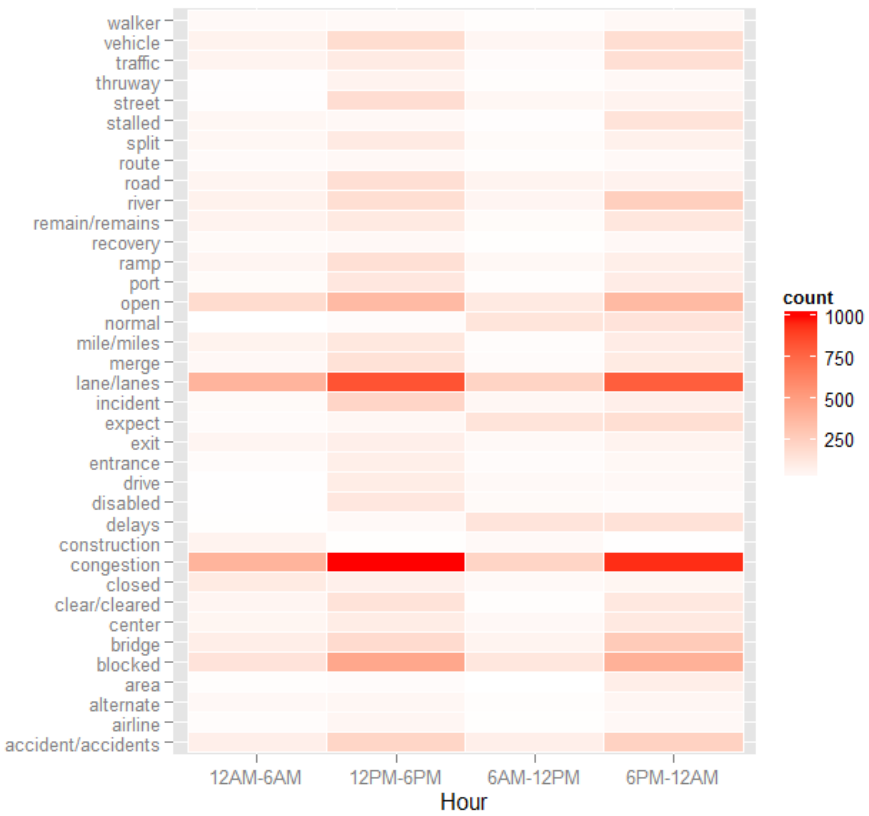


(a) BR_Traffic

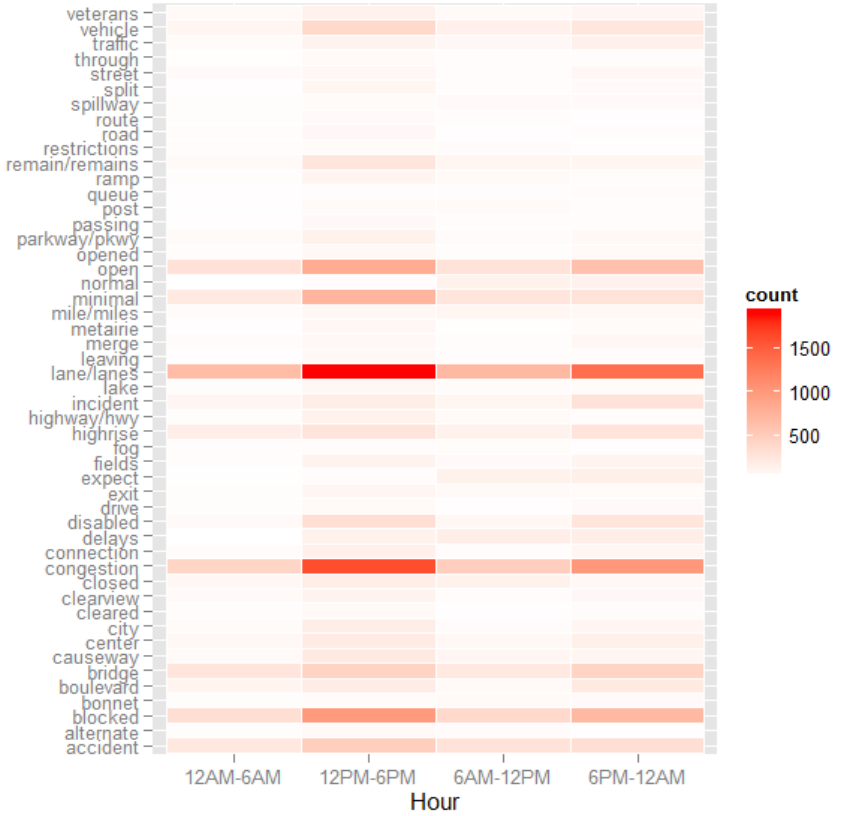


(b) NOLA_Traffic

FIGURE 3 Frequency of terms

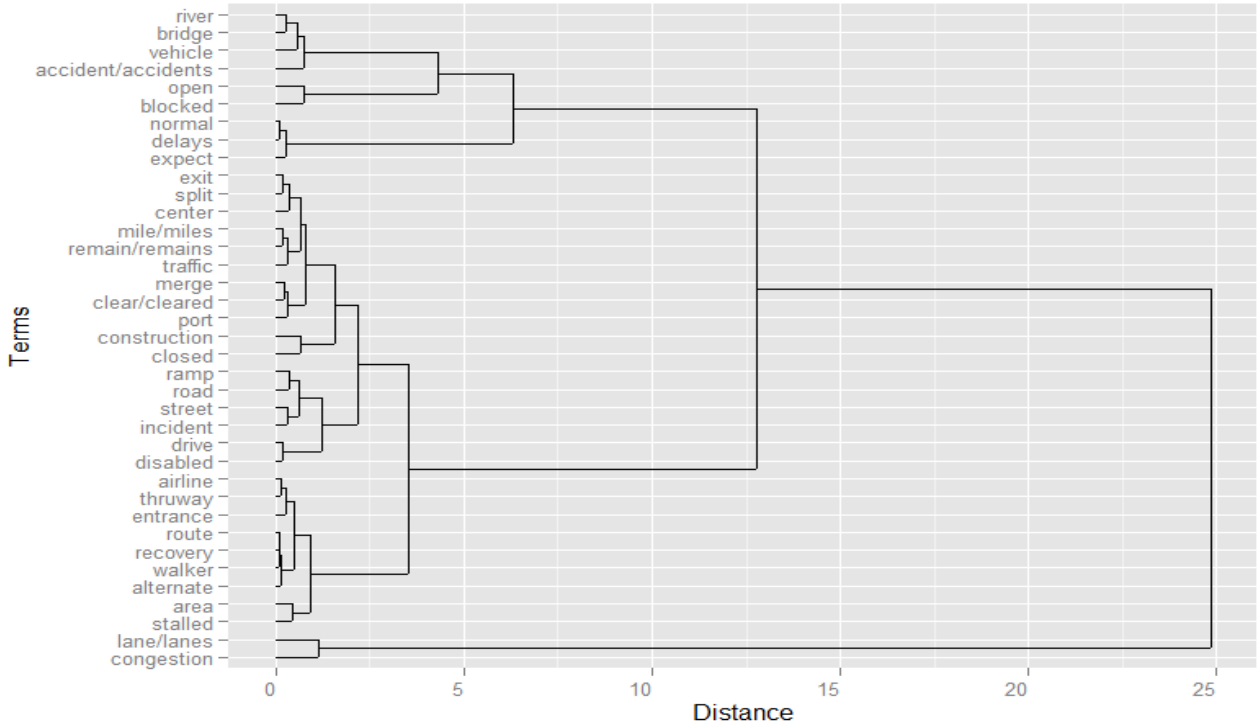


(a) BR_Traffic

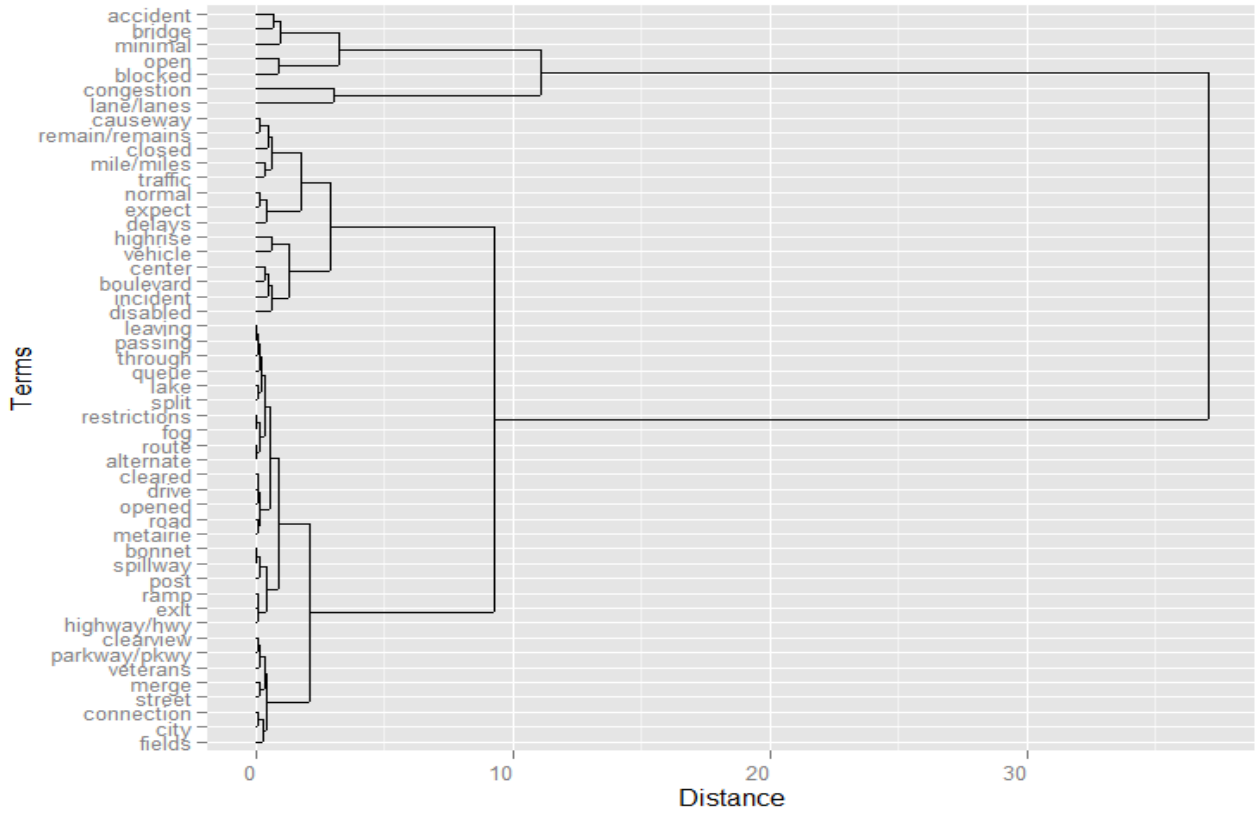


(b) NOLA_Traffic

FIGURE 4 Heat map of terms per corpus



(a) BR_Traffic



(b) NOLA_Traffic

FIGURE 5 Hierarchical clustering dendrogram

Sentiment Analysis

Sentiments are central to almost all human activities and are key influencers of our behaviors. Most human beliefs and perceptions of humankind depends on how others see and evaluate the world. For this reason, people often seek out the sentiments of others in order to make a better decision. This is not only true for individuals but also true for various programs and organizations. Opinions and related concepts such as sentiments, evaluations, attitudes, and emotions are the subjects of study of sentiment analysis. Figure 3 illustrates the flowchart of the sentiment analysis procedure conducted in this study.

It is important to note that the sentiment lexicons have domain-specific sentiment values; therefor sentiment classification performance of a given text may vary according to the calculation process of the sentiment for that text. Various sentiment lexicons with different format and research focus have been developed to aid the classification of positive and negative annotations in the mining-ready texts. We noticed both similarity and diversification while comparing the listed words and their ratings. Constructing a domain-specific sentiment lexicon is essential to tackle the classification problem of sentiment analysis. The researchers of this study are currently developing a sentiment lexicon appropriate for transportation related tweets. This work remains a prospective topic for future research. We used a list of positive and negative sentiment words in English to perform the sentiment analysis on the tweets. The list used was the list compiled by Hu and Liu in 2004 [22].

For example, it will be interesting to mine the Twitter data related to "@NOLA_Traffic" and "#NOLA_Traffic" to understand the sentiment of the New Orleans roadway users. A function named "score.sentiment", introduced by Breen, was used to produce the score count of each tweet [23]. The researchers of this project modified this function. This function will mine each tweet by using the positive and negative word lexicons and produce a positive, negative or zero score. A tweet with a "+2" score means that this particular tweet has two positive words by mentioning or hashtagging "NOLA_Traffic". A tweet with a negative score indicates the negative words used in a particular tweet.

We collected tweets related to four search terms: workzone, redlight camera, seatbelt, and pavement marks. The terms workzone and pavement marks show higher trends in positive scores while terms like redlight camera, and seatbelt process more inclination

towards negative scores. Figure 6 illustrates these scores.

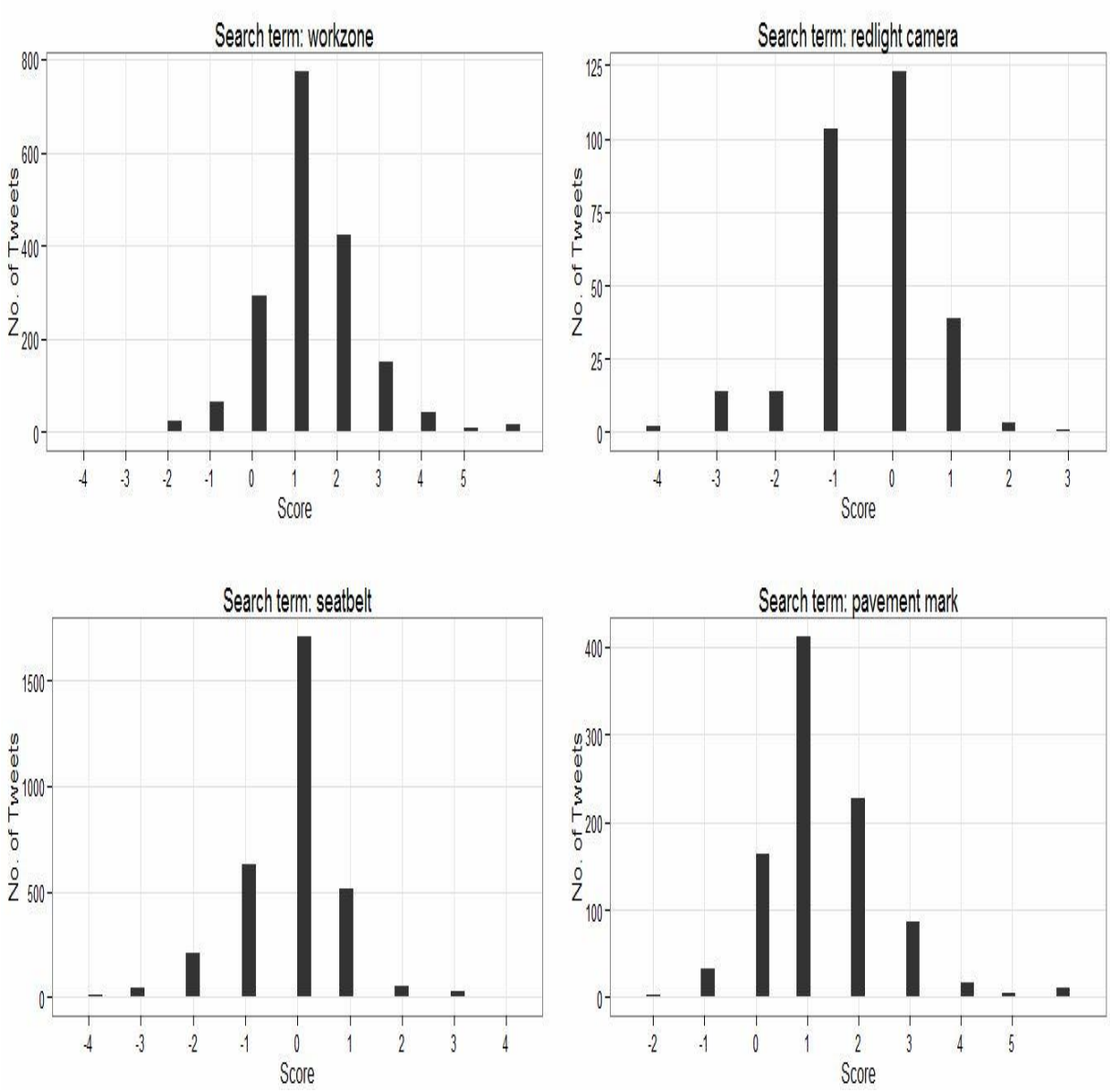


FIGURE 6 Sentiment scores of four countermeasures

We also collected tweets related to two DOTD Twitter handles: BR_Traffic, and NOLA_Traffic. Figure 7 illustrates these scores. As both of the DOTD Twitter handles tweet regarding congestion, block, accident/crash (mostly negative terms), the inclination towards negative scores are heavier.

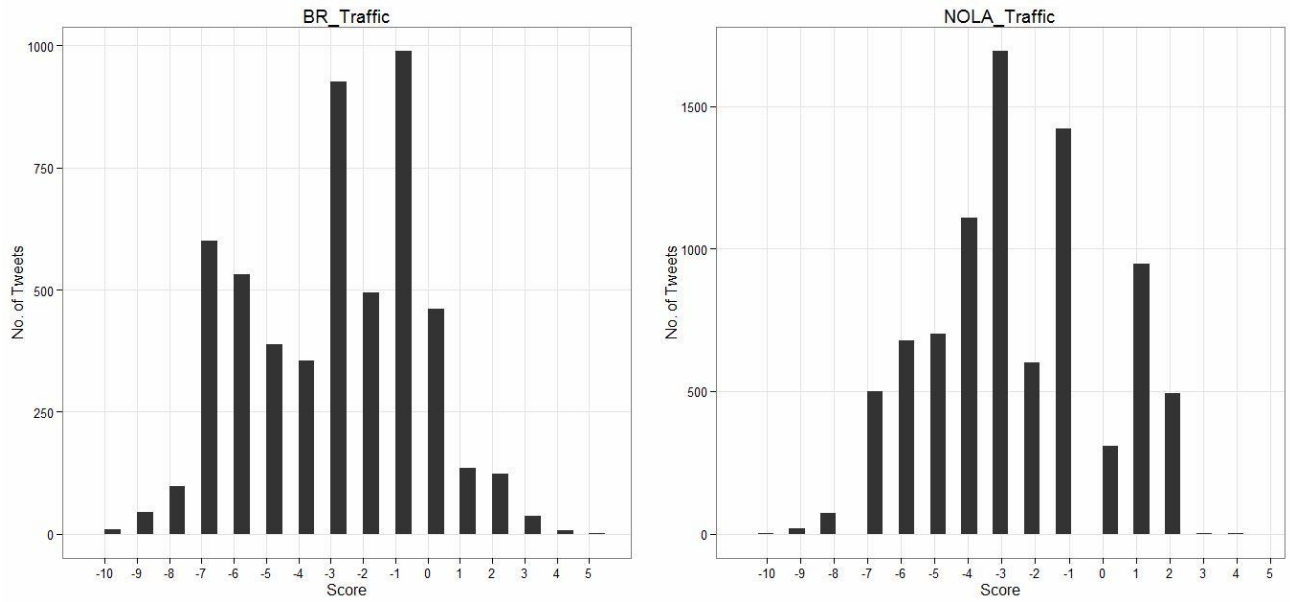


FIGURE 7 Sentiment scores of two DOTD Twitter handles

DISCUSSIONS OF RESULTS

This project performed Twitter mining on the collected tweets from the DOTD official Twitter handles. We performed the analysis in two specific ways: 1) text mining on the generated tweets to discover knowledge patterns, and 2) sentiment analysis on specific traffic countermeasures and DOTD Twitter handle tweets. The text mining results show an inclination towards specific terms. The terms indicate that DOTD official tweets are mostly information related and circulation of these tweets (in real-time through mobile apps) will be helpful in reducing crashes and congestion. The sentiment analysis on four specific countermeasures showed mixed sentiments. Countermeasures effective with traffic fines show more inclination towards negative scores.

CONCLUSIONS

Twitter has the potential to improve public service. Evaluation of social media inputs is thus necessary. At the same time, it needs to be inquired whether this potential has emerged or not. In this context, we have presented a general picture regarding the Twitter usage by DOTD transportation authorities and investigated the purposes of Twitter usage. Our analysis has revealed that social media usage in governmental information sharing is significant. It also enables the roles of citizens in co-producing/retweeting public information services under circumstances of extreme weather. A benefit-cost analysis would clearly quantify the impact of social media usage for this particular case. This study demonstrates that text mining retrieved knowledge from a DOTD's official tweets. The frequent term analysis from both of the Twitter handles is similar. The tweets are mostly related to terms like 'congestion', 'blocked', 'lane/lanes', 'accident' and 'open'. Real-time usage of these tweets are beneficial if they are supplied to people in real-time. This study has three particular contributions: 1) it developed a framework of data collection related to transportation information tweets, 2) it developed a text-mining framework to extract knowledge for integration in various perspectives, and 3) it performed sentiment analysis on the DOTD official tweets and on four specific countermeasures. Future research can be directed towards several scopes from the current study: economic and environmental impact analysis of real-time information sharing, sentiment analysis of the Louisiana Twitter users on government public services, and impact of congestion/blockage information shared by DOTD Twitter handles by using the Google Map application-programming interface (API). Finally, we can say that social innovation through information sharing is creating an environment in which governments and citizens can work together and may fundamentally modify the centrality of governments in making proactive policies.

RECOMMENDATIONS

This project develops unique tools for understanding people's sentiments on specific terms related to transportation safety. Transport authorities can utilize the developed algorithm and the senti-lexicon for further improvement to make real-time information feeds or alerts available through mobile apps.

ACRONYMS, ABBREVIATIONS, AND SYMBOLS

API	Application Programming Interface
CART	Classification and Regression Tree
DOT	Department of Transportation
DOTD	Louisiana Department of Transportation and Development
FHWA	Federal Highway Administration
Hwy	Highway
LTRC	Louisiana Transportation Research Center
MPO	Metropolitan Planning Organization
U.S. Hwy	United States Highways
OAuth	Open standard for Authorization
VMT	Vehicle Mile Traveled
vpd	Vehicles Per Day

REFERENCES

1. Mooney, R., and Bunescu, R. Mining knowledge from text using information extraction. ACM SIGKDD Newsletter 7, pp. 3–10, 2005.
2. Meystre, S., Savova, G., Kipper-Schuler, K., and Hurdle, J. Extracting information from textual documents in the electronic health record: a review of recent research. Yearbook of Medical Information, pp. 128–144, 2008.
3. Pennacchiotti, M., and Gurumurthy, S. Investigating topic models for social media user recommendation. In Proceedings of the 20th international conference companion on World Wide Web WWW '11, pp. 101–102, New York, 2011.
4. Lin, C., and He, Y. Joint sentiment/topic model for sentiment analysis. Proceedings of the 18th ACM conference on Information and knowledge management CIKM '09, pp. 375–384, New York, 2009.
5. Duan, J., and Zeng, J. Web objectionable text content detection using topic modeling technique. Expert Systems with Applications, 40, pp. 6094–6104, 2013.
6. Martinez-Romo, J., and Araujo, L. Detecting malicious tweets in trending topics using a statistical analysis of language. Expert Systems with Applications, 40, pp. 2992–3000, 2013.
7. Waters, R. and Jamal, J. Tweet, tweet, tweet: A content analysis of nonprofit organizations' Twitter updates. Public Relations Review 37, pp. 321– 324, 2011.
8. Lee, C. Unsupervised and supervised learning to evaluate event relatedness based on content mining from social-media streams. Expert Systems with Applications, 39, 13338–13356, 2012.
9. Panagiotopoulos , P., Bigdeli, A., and Sams, S. Citizen–government collaboration on social media: The case of Twitter in the 2011 riots in England. Government Information Quarterly, in press, 2014.
10. Sobaci, M., and Karkin, N. The use of twitter by mayors in Turkey: Tweets for better public services? Government Information Quarterly, Vol. 30, pp. 417–425, 2013.
11. Chatfield, A., Scholl, H., and Brajawidagda, U. Tsunami early warnings via Twitter in government: Net-savvy citizens' co-production of time-critical public information services. Government Information Quarterly, Vol. 30, pp. 377–386, 2013.
12. Hong, S. Online news on Twitter: Newspapers' social media adoption and their online readership. Information Economics and Policy, Vol. 24, pp. 69–74, 2012.
13. Bollen, J., Mao, H., and Zeng, X. Twitter mood predicts the stock market. Journal of Computer Science, Vol. 2 (1), pp. 1–8, 2011.

14. Sakaki, T., Okazaki, M., and Matsuo, Y. Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web. WWW'10. ACM, New York, pp. 851–860, 2010.
15. Culotta, A. Towards detecting influenza epidemics by analyzing twitter messages. Proceedings of the First Workshop on Social Media Analytics.SOMA'10. ACM, New York, pp. 115–122, 2010.
16. Borondo, J., Morales, A., Losada, J., and Benito, R. Characterizing and modeling an electoral campaign in the context of twitter: 2011 Spanish presidential election as a case study. Chaos, Vol. 22 (2), 2012.
17. Bibliography of social media research. https://dl.dropboxusercontent.com/u/13642258/socialmedia2.html. Last retrieved on June 17, 2015.
18. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (2008). Advances in Knowledge Discovery and Data Mining, MIT Press, 1996.
19. Developer website for Twitter. https://dev.twitter.com/docs/auth/oauth. Accessed July 26, 2014.
20. I. Feinerer, K. Hornik, and D. Meyer (2008). Text Mining Infrastructure in R. Journal of Statistical Software, Vol. 25(5), pp. 1-54.
21. Listing of DOTD Official tweets. https://dl.dropboxusercontent.com/u/13642258/DOTD_Official_TW14.html. Last retrieved on June 17, 2015.
22. Liu, B. (2012) Sentiment Analysis and Opinion Mining (Synthesis Lectures on Human Language Technologies). Morgan & Claypool Publishers, Vermont, Australia.
23. Breen, J.R. (2014) Tutorial on Twitter Text Mining. https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial-201107. Last retrieved on June 17, 2015.

APPENDIX A

R codes

```
### TWEET COLLECTION

library(twitteR)
require(twitteR)
require(ROAuth)

requestURL <- "https://api.twitter.com/oauth/request_token"
accessURL <- "https://api.twitter.com/oauth/access_token"
authURL <- "https://api.twitter.com/oauth/authorize"
consumerKey <- "my_consumer_key"
consumerSecret <- "my_consumer_secret"
twitCred <- OAuthFactory$new(consumerKey=consumerKey,
                             consumerSecret=consumerSecret,
                             requestURL=requestURL,
                             accessURL=accessURL,
                             authURL=authURL)


setwd("my_folder")
download.file(url="http://curl.haxx.se/ca/cacert.pem",
destfile="cacert.pem")
twitCred$handshake(cainfo="cacert.pem")

registerTwitterOAuth(twitCred)
save(list="twitCred", file="twitteR_credentials")
load("twitteR_credentials")
registerTwitterOAuth(twitCred)

nola_tweets = userTimeline("NOLA_Traffic", n=3200)
nola_tweets1 <- twListToDF(nola_tweets)
write.csv(nola_tweets1, " nola_tweets.csv")


br_tweets = userTimeline("BR_Traffic", n=3200)
br_tweets1 <- twListToDF(br_tweets)
write.csv(nola_tweets1, " br_tweets.csv")

seatbelt = searchTwitter("seatbelt", cainfo="cacert.pem", lang=
"en", n=3200)
seatbelt1 <- twListToDF(seatbelt)
write.csv(seatbelt1, "seatbelt.csv")
```

```
redlight = searchTwitter("redlight camera", cainfo="cacert.pem",
lang= "en", n=3200)
redlight1 <- twListToDF(redlight)
write.csv(redlight1, "redlight.csv")

workzone = searchTwitter("workzone ", cainfo="cacert.pem", lang=
"en", n=3200)
workzone1 <- twListToDF(workzone)
write.csv(workzone1, "workzone.csv")

pavement_mark = searchTwitter("pavement marking",
cainfo="cacert.pem", lang= "en", n=3200)
pavement_mark1 <- twListToDF(pavement_mark)
write.csv(pavement_mark1, "pavement_mark.csv")



### TEXT MINING ON COLLECTED TWEETS

### Used R packages
library(ggplot2)
library(scales)
library(lubridate)
library(gridExtra)
library(tm)
library(reshape)



#### BR_Tweets

setwd("data_folder")
tweets <- read.csv("BR_TrafficALL_trun.csv" )
head(tweets)
dim(tweets)
table(tweets$Month_Cat)
table(tweets$Hour_Cat)

##### MONTH

all2<- data.frame(Month = unique(tweets$Month_Cat),
Tweets = tapply(tweets$text, tweets$Month_Cat, paste, collapse =
' '))
names(all2)
dim(all2)

#### HOUR
```

```
all21<- data.frame(Hour = unique(tweets$Hour_Cat),
Tweets = tapply(tweets$text, tweets$Hour_Cat, paste, collapse =
' '))
names(all21)
dim(all21)


mydata.corpus <- Corpus(VectorSource(all2$Tweets))
mydata.corpus <- tm_map(mydata.corpus, tolower)
mydata.corpus <- tm_map(mydata.corpus, removePunctuation,
preserve_intra_word_dashes=TRUE)
my_stopwords <- c(stopwords('german'),"the", "due", "are",
"not", "for", "this", "and",  "that", "there", "new", "near",
"beyond", "time", "from", "been", "both", "than",
"has","now", "until", "all", "use", "two", "ave", "blvd",
"east", "between", "ccc", "end", "have", "avenue", "before", "i-
us", "i-e", "i-i-", "ames", "belle", "gen", "okeefe", "one",
"just", "mac", "being", "i-i-", "tchoupitoulas", "williams",
"left", "right","bonnabel", "tulane", "west", "franklin",
"lafayette","louisia", "orleans", "pontchartrain", "paris" ,
"still", "off", "over", "only", "north", "past",
"twin", "while", "menteur" , "i-w", "loyola", "manchac" ,
"manhattan" , "south", "arthur", "barataria" , "bayou" ,
"bernard", "carre" ,"carrollton", "crescent" , "gaulle" ,
"general" , "harvey", "i-e","i-i-","i-us" , "must", "more",
"work","read", "poydras",  "reached", "morrison", "louisa",
"earhart", "elysian", "charles", "claiborne", "chef", "wisner" ,
"mph", "three","info", "canal", "camp", "la-", "approximately",
"essen", "acadian",  "perkins", "dalrymple", "chippewa",
"baton", "rouge", "amp", "access", "approaching", "highland",
"washington", "sherwood", "siegen", "prairieville",
"mississippi", "mrb", "livingston", "louise", "i-i-", "i-e","i-
w","florida", "government", "forest", "friday",  "drusilla",
"capitol", "bluebonnet")

mydata.corpus <- tm_map(mydata.corpus, removeWords,
my_stopwords)
mydata.corpus <- tm_map(mydata.corpus, removeNumbers)

mydata.dtm <- TermDocumentMatrix(mydata.corpus)
inspect(mydata.dtm[1:4,1:4])

DTM <- DocumentTermMatrix(mydata.corpus)
inspect(DTM[1:4,1:4])

findFreqTerms(mydata.dtm, lowfreq=10)
findAssocs(mydata.dtm, 'congestion', 0.7)
```

```
mydata.dtm2 <- removeSparseTerms(mydata.dtm, sparse=0.01)
inspect(mydata.dtm2[1:4,1:4])

library(slam)
TDM.dense <- as.matrix(mydata.dtm2)
TDM.dense
object.size(mydata.dtm2)
object.size(TDM.dense)

#### NOLA_Tweets

setwd("data_folder")
tweets <- read.csv("NOLA_TrafficALL_trun.csv" )
head(tweets)
dim(tweets)
table(tweets$Month_Cat)
table(tweets$Hour_Cat)

##### MONTH

all2<- data.frame(Month = unique(tweets$Month_Cat),
Tweets = tapply(tweets$text, tweets$Month_Cat, paste, collapse =
' '))
names(all2)
dim(all2)

#### HOUR
all21<- data.frame(Hour = unique(tweets$Hour_Cat),
Tweets = tapply(tweets$text, tweets$Hour_Cat, paste, collapse =
' '))
names(all21)
dim(all21)


mydata.corpus <- Corpus(VectorSource(all2$Tweets))
mydata.corpus <- tm_map(mydata.corpus, tolower)
mydata.corpus <- tm_map(mydata.corpus, removePunctuation,
preserve_intra_word_dashes=TRUE)
my_stopwords <- c(stopwords('german'),"the", "due", "are",
"not", "for", "this", "and",  "that", "there", "new", "near",
"beyond", "time", "from", "been", "both", "than",
"has","now", "until", "all", "use", "two", "ave", "blvd",
"east", "between", "ccc", "end", "have", "avenue", "before", "i-
us", "i-e", "i-i-", "ames", "belle", "gen", "okeefe", "one",
"just", "mac", "being", "i-i-", "tchoupitoulas", "williams",
```

```
"left", "right","bonnabel", "tulane", "west", "franklin",
"lafayette","louisia", "orleans", "pontchartrain", "paris" ,
"still", "off", "over", "only", "north", "past",
"twin", "while", "menteur" , "i-w", "loyola", "manchac" ,
"manhattan" , "south", "arthur", "barataria" , "bayou" ,
"bernard", "carre" ,"carrollton", "crescent" , "gaulle" ,
"general" , "harvey", "i-e","i-i-","i-us" , "must", "more",
"work","read", "poydras",  "reached", "morrison", "louisa",
"earhart", "elysian", "charles", "claiborne", "chef", "wisner" ,
"mph", "three","info", "canal", "camp", "la-", "approximately",
"essen", "acadian",  "perkins", "dalrymple", "chippewa",
"baton", "rouge", "amp", "access", "approaching", "highland",
"washington", "sherwood", "siegen", "prairieville",
"mississippi", "mrb", "livingston", "louise", "i-i-", "i-e","i-
w","florida", "government", "forest", "friday",  "drusilla",
"capitol", "bluebonnet")

mydata.corpus <- tm_map(mydata.corpus, removeWords,
my_stopwords)
mydata.corpus <- tm_map(mydata.corpus, removeNumbers)

mydata.dtm <- TermDocumentMatrix(mydata.corpus)
inspect(mydata.dtm[1:4,1:4])

DTM <- DocumentTermMatrix(mydata.corpus)
inspect(DTM[1:4,1:4])

findFreqTerms(mydata.dtm, lowfreq=10)
findAssocs(mydata.dtm, 'congestion', 0.7)


mydata.dtm2 <- removeSparseTerms(mydata.dtm, sparse=0.01)
inspect(mydata.dtm2[1:4,1:4])

library(slam)
TDM.dense <- as.matrix(mydata.dtm2)
TDM.dense
object.size(mydata.dtm2)
object.size(TDM.dense)


### SENTIMENT ANALYSIS

hu.liu.pos = scan('positive-words.txt', what='character',
comment.char=';')
hu.liu.neg = scan('negative-words.txt', what='character',
comment.char=';')
```

```
pos.words = c(hu.liu.pos, 'upgrade')
neg.words = c(hu.liu.neg, 'wtf', 'wait', 'waiting', 'epicfail',
'mechanical')

neglist <- c('congestion', 'blocked', 'accident','delays',
'closed', 'stalled','incident')
poslist <- c('open', 'minimal', 'recovery', 'cleared', 'clear')


score.sentiment = function(sentences, pos.words, neg.words,
.progress='none')
{
require(plyr)
require(stringr)
# we got a vector of sentences. plyr will handle a list or a
vector as an "l" for us
# we want a simple array of scores back, so we use "l" + "a" +
"ply" = laply:
scores = laply(sentences, function(sentence, pos.words,
neg.words) {
# clean up sentences with R's regex-driven global substitute,
gsub():
sentence = gsub('[[:punct:]]', '', sentence)
sentence = gsub('[[:cntrl:]]', '', sentence)
sentence = gsub('\\d+', '', sentence)
# and convert to lower case:
sentence = tolower(sentence)
# split into words. str_split is in the stringr package
word.list = str_split(sentence, '\\s+')
# sometimes a list() is one level of hierarchy too much
words = unlist(word.list)
# compare our words to the dictionaries of positive & negative
terms
pos.matches = match(words, pos.words)
neg.matches = match(words, neg.words)
# match() returns the position of the matched term or NA
# we just want a TRUE/FALSE:
pos.matches = !is.na(pos.matches)
neg.matches = !is.na(neg.matches)
pos.matches1 = match(words, poslist)
neg.matches1 = match(words, neglist)
pos.matches1 = !is.na(pos.matches1)
neg.matches1 = !is.na(neg.matches1)

score = sum(pos.matches) + 2*sum(pos.matches1)-
(sum(neg.matches)+2*sum(neg.matches1))
return(score)
```

```
}, pos.words, neg.words, .progress=.progress )
scores.df = data.frame(score=scores, text=sentences)
return(scores.df)
}


a1.text <- read.csv("workzone.csv" )
a1.scores = score.sentiment(a1.text$text, pos.words, neg.words,
.progress='text')
plot1 <- qplot(a1.scores$score)
plot11 <- plot1 +xlab("Score") + ylab("No. of
Tweets")+theme_bw()+ scale_x_continuous(breaks=-5:5)+
labs(title = "Search term: workzone")


a2.text <- read.csv("redlight.csv" )
a2.scores = score.sentiment(a2.text$text, pos.words, neg.words,
.progress='text')
plot2 <- qplot(a2.scores$score)
plot21 <- plot2 +xlab("Score") + ylab("No. of
Tweets")+theme_bw()+ scale_x_continuous(breaks=-5:5)+
labs(title = "Search term: redlight camera")

a3.text <- read.csv("seatbelt.csv")
a3.scores = score.sentiment(a3.text$text, pos.words, neg.words,
.progress='text')
plot3 <- qplot(a3.scores$score)
plot31 <- plot3 +xlab("Score") + ylab("No. of
Tweets")+theme_bw()+ scale_x_continuous(breaks=-5:5)+
labs(title = "Search term: seatbelt")

a4.text <- read.csv("pavement_mark.csv")
a4.scores = score.sentiment(a4.text$text, pos.words, neg.words,
.progress='text')
plot4 <- qplot(a4.scores$score)
plot41 <- plot4 +xlab("Score") + ylab("No. of
Tweets")+theme_bw()+ scale_x_continuous(breaks=-5:5)+
labs(title = "Search term: pavement mark")

library(gridExtra)
grid.arrange(plot11, plot21, plot31, plot41, ncol=2)


a5.text <- read.csv("BR_TrafficALL_April24_trun.csv")
a5.scores = score.sentiment(a5.text$text, pos.words, neg.words,
.progress='text')
plot5 <- qplot(a5.scores$score)
```

```
plot51 <- plot5 +xlab("Score") + ylab("No. of
Tweets")+theme_bw()+ scale_x_continuous(breaks=-10:5)+
labs(title = "BR_Traffic")

a6.text <- read.csv("NOLA_TrafficALL_April24trun.csv")
a6.scores = score.sentiment(a6.text$text, pos.words, neg.words,
.progress='text')
plot6 <- qplot(a6.scores$score)
plot61 <- plot6 +xlab("Score") + ylab("No. of
Tweets")+theme_bw()+ scale_x_continuous(breaks=-10:5)+
labs(title = "NOLA_Traffic")

library(gridExtra)
grid.arrange(plot51, plot61, ncol=2)
```

APPENDIX B

Hierarchical Clustering: Theory

A hierarchical clustering, a widely used data analysis tool, works by building a binary tree of the data that successively merges similar groups of points. Two methods are dominant based on the decomposition: agglomerative and divisive hierarchical clustering. If agglomerative clustering acts as a bottom-up method, then we can consider decisive clustering as a bottom-down method. One can visualize the results of the hierarchical clustering with the use of a dendrogram, a valued tree. It is an n-tree where each node is associated with a height satisfying the condition:

$$h(A) \leq h(B) \Leftrightarrow A \subseteq B$$

If A ∩ B= φ, this condition is true for all data points of A and B. Here, h (A) and h (B) denote the heights of A and B respectively.

We adopt different methods in hierarchical clustering. Ward's method forms the partition P_n, P_{n-1}, ……, P_1 in such a way that minimizes the loss of information associated with each of the merging. Usually, we quantify the loss of information in terms of an error sum of squares (SSE). Ward's method is therefore known as the 'minimum variance' method. The distance between the two clusters, A and B, is how much the sum of squares increases while merging:

$$dist(A,B) = \sum_{i \in A \cup B} \left\| \vec{y}_i - \vec{m}_{A \cup B} \right\|^2 - \sum_{i \in A} \left\| \vec{y}_i - \vec{m}_A \right\|^2 - \sum_{i \in B} \left\| \vec{y}_i - \vec{m}_B \right\|^2 = \frac{c_A c_B}{c_A + c_B} \left\| \vec{m}_A - \vec{m}_B \right\|^2 \quad (1)$$

where, $\vec{m}_j$ is the center of cluster j and c_j is the number of the points in that cluster.