# A Novel Transit Rider Satisfaction Metric: Rider Sentiments Measured from Online Social Media Data

*Craig Collins, Samiul Hasan, Satish V. Ukkusuri, Ph.D.*
*Purdue University*

## Abstract

*The goal of this paper is to demonstrate the use of an innovative social media-based data source, Twitter, to evaluate transit rider satisfaction. Transit authorities have access to vast amounts of performance metrics that measure ridership, timeliness, efficiency, safety, cleanliness, and service, to name a few. These performance metrics, however, are generally one-sided; they represent the interests of the business and are not customer-based. This paper recognizes the limitations of standard performance metrics and attempts to gauge transit rider sentiments by measuring Twitter feeds. Sentiment analysis is used to classify a population of rider sentiments over a period of time. Conclusions are drawn from totals of positive and negative sentiments, normalized average sentiments, and the total number of Tweets collected over a time period.*

## Introduction

With the advent of social media, people are able to express their opinions on a subject instantaneously. Researchers are beginning to "mine" these opinions from social media outlets (i.e., Twitter, Facebook) to form general public perceptions, or sentiments, on a number of subjects. Sentiment analysis, or opinion mining, is the examination of a text through software to understand the positive or negative connotations surrounding it. Sentiment analysis can assist companies in determining how a brand is perceived in relation to value and quality. Subtasks of sentiment analysis include determining subjectivity, polarity (positive or negative), and degree of polarity, and classifying the subject matter and author. Sentiment analysis, through the monitoring of social media, can change the way transit agencies measure rider satisfaction.

For trips with similar distances, transit agencies are in direct competition with automobile services as well as pedestrian modes such as walking and bicycling. Commuters are likely to choose modes that maximize their utility and provide the most satisfaction (Andreassen 1995). Studies have been conducted that show commuters are more likely to choose modes that are more economical, comfortable, work-friendly, punctual, and safe (Hanna and Drea 1998). Thevathasan and Balachandran (2007) found that transit passengers consider improvements to safety and stations, improvements in facilities, and cleanliness to be important. Unfortunately for transit agencies, there is often a lack of compatibility between passenger needs and management perception of those needs. Management runs the risk of misallocating scarce resources and not being aware of growing passenger dissatisfaction with transit services that are measured by performance metrics (Koushki et al. 2003).

Performance metrics are constructed to encourage performance improvement, effectiveness, efficiency, and appropriate levels of internal controls. Traditionally, performance metrics are measured through quantities related to ridership, timeliness, efficiency, safety, cleanliness, service, and courtesy (Chicago Transit Authority 2011; Metropolitan Transit Authority 2011). The *Transit Capacity and Quality of Service Manual (TCQSM)* (1999) recognizes that there is a distinct difference between transit agency and passenger point of view for quality of service. *TCQSM* indicates service coverage, hours of service, amenities, safety, and travel time as relevant to passengers, whereas annual ridership, vehicles operated in maximum service, passenger miles/revenue hour, and vehicle operating expenses are relevant to transit agencies. An analysis of Portland, Oregon's, local transit provider (TriMet) shows how technology can be used to develop an archival system that directly relates bus transit performance to performance indicators through the use of its bus dispatch system (Bertiniand and El-Geneidy 2003).

With the ever-growing complexities of urban transit and decreasing of federal and state funding, many transit authorities may not have the means to monitor zones with high amounts of activity. Through the use of online social media, commuters can voice their concerns in real time about current conditions in service, safety, sanitary conditions, etc. Outside of surveys and focus groups, there are few ways to gauge customer satisfaction, let alone in a practical and timely fashion. The purpose of this study was to show a proof of concept using an emerging data source, such as Twitter, and conduct a sentiment analysis using SentiStrength (Thelwall et al. 2010) to evaluate transit rider satisfaction. This study is intended as a pilot test and demonstrates the usefulness of social media data for measuring rider perception.

The data used in this work were obtained from the riders of the rapid transit system of the Chicago Transit Authority (CTA). CTA is an ideal setting for analysis because of its large volume of riders who depend on the its rapid transit system for commuting to work and for navigating the city of Chicago. CTA operates the nation's second largest public transportation system and serves a population of 3.8 million (Chicago Transit Authority 2011). In addition to serving the city of Chicago, 40 suburbs also rely on the CTA's transit system. On the rapid transit system, CTA's 1,200 rail cars operate over 8 lines (Blue, Brown, Green, Orange, Pink, Purple, Red, and Yellow) and 224.1 miles of track. CTA trains make about 2,145 trips each day and serve 143 stations. On average 641,261 riders rode the CTA rail system per weekday in 2010. CTA's rapid transit system is referred to as the "L," having gained its nickname from large parts of the system being elevated, although segments of the network are underground, at grade level, and open cut. CTA has rail service in proximity to two airports (O'Hare International and Midway), professional sporting arenas (Wrigley Field, Soldier Field, United Center), city government buildings, and many recreational sites (parks, zoos, waterfront).

CTA possesses a strong need for sentiment analysis because of its recent fiscal restraints. The State of Illinois has once again cut funds to CTA due to a lack of funds at the state level. In 2010, CTA cut 9 percent of its rail service and laid off nearly 1,100 employees. It has been forced to use much of its capital funds to maintain day-to-day operations. Sentiment analysis can provide customer feedback on fare increases, services, and provide a means of monitoring safety due to a lack of personnel.

To analyze rider sentiment of CTAs "L" system, a large population of data is needed. In this work, Twitter texts, or "tweets," were collected and analyzed to quantify and compare performance. Twitter is a real-time information network that connects users to the latest information about what they find interesting. Users share opinions and general statements through "tweets," short texts of 140 characters or less. Given the wide prevalence of Twitter data, it provides rich data sources to measure the sentiment of transportation systems. Twitter currently receives around 140 million tweets on average daily.

Twitter's most appealing feature is that information can be viewed and analyzed using specialized crawlers that collect the data in real time. Searches by keyword and proximity to a location are only a few of the ways Twitter allows for searches of relevant tweets. Twitter provides a time-stamp to tweets, which is beneficial in determining the relevance of a text and organizing sentiments over a time frame. Twitter is popular for its accessibility and mobility because it allows its users to upload and view comments via laptop, tablet, and smart phone. While Twitter can provide a vast amount of information on a given topic, not all the data returned are relevant. Keyword searches can bring up material unrelated to the desired topic. For example, in our research, one of the keywords used was "red line." Red line is not only associated with a particular CTA "L"; many tweets were retrieved with views on a 1998 movie called "The Thin Red Line." Because tweets can have a maximum of 140 characters, users must often abbreviate words and phrases, and the overall message of these texts can be lost or misunderstood. It was observed that roughly 25 percent of the data collected on a topic is relevant and opinion-based. Other data come in the form of Twitter applications such as FourSquare, company advertisements, and factual statements by users.

## Related Works

With the advent of the Internet and social media, research has intensified in the field of sentiment analysis. Sentiment classification subtasks revolve around determining subjectivity, polarity (positive or negative), and the degree of polarity, and classifying the subject matter and author. Sentiment analysis is based on two techniques: the use of linguistic resources, where each word is assigned a value, and machine learning techniques, which use counting methods to determine the sentiment of a body of text. Machine learning techniques are concerned with the construction of algorithms to allow computers to model behaviors based on available data. This literature review focuses on past studies that have used sentiment analysis over a wide range of applications.

Thomas et al. (2006) determined whether discussions among speakers in U.S. Congressional debates were in support of or opposition to the topic being discussed. Niu et al. (2005) evaluated the problem of detecting the presence of a clinical outcome in medical texts, and, when an outcome was found, determining whether it was positive, negative, or neutral, as shown in the following examples.

Text summarization can be used in information filtering. It is important for data miners to be able to find relevant information that suits their needs. Kim and Hovy (2004) presented a system that, given a topic, automatically finds the people who hold opinions about that topic and the sentiment of each opinion. Their system contained a module for determining word sentiment and another for combining sentiments within a sentence.

The research of sentiment analysis has been conducted on many applications and is just the beginning in understanding how sentiment analysis can be applied to a variety of fields. Future applications include Web mining for consumer opinion summarization, business and government intelligence analysis, and improving text analysis applications such as information retrieval, question answering, and text summarization (Finn et al. 2002). Sentiment analysis uses real-time data from blog posts, reviews and opinions posted on Web sites, and status messages and comments posted in social media that provide feedback on current happenings in relation to a topic. Examples of sentiment analysis include determining sentiment characteristics in Web blogs (Pang and Lee 2008), polarity detection in Congressional debates (Thomas et al. 2006), summarizing movie reviews (Turney 2001; Zhuang et al. 2006), spam detection in product reviews (Jindal and Liu 2007; Jindal and Liu 2008), and predicting the stock market (Bollen and Mao 2011).

## Methodology

### *SentiStrength: A Sentiment Strength Detection Algorithm*

In this study, we used SentiStrength (SentiStrength 2011; Thelwall et al. 2010), a machine learning program that detects the sentiment value of a short text, for analyzing the sentiments of rider tweets about the transit system. Users input one or more texts, and the general strength of the sentiment behind each text was quantified. Output from the software can document the average negative and positive sentiment or the minimum negative or maximum positive sentiment. SentiStrength uses machine learning techniques to conduct sentiment analysis. The core of its algorithms is the sentiment word strength list (Thelwall et al. 2010), which can be updated with new words relevant to a specific topic. These words can then be updated, or optimized, allowing SentiStrength to better judge sentiment polarity.

SentiStrength contains an original collection of 298 positive terms and 465 negative terms classified with relative sentiment strength. Ratings of 1 to 5 indicate a positive sentiment and -1 to -5 indicate a negative sentiment, with -5 and 5 being the maximum negative and positive sentiments that can be attained. The word list was combined with the AFINN word strength list to bring the total word strength list to 2,509 terms. AFINN is an English word list with 2,477 words constructed by Finn Årup Nielsen for sentiment analysis of Twitter messages (Hansen et al. 2010). Each word is rated by a valence value from -5 to +5, excluding 0. AFINN was ideal because it uses a numerical system identical to SentiStrength, placing word strength values between 1 and 5 for either positive or negative sentiment strength. Allowing the user to update the word strength list is important because, initially, the list is relatively small, in comparison to every other imaginable word. Also, allowing the user to update the word strength list enables SentiStrength to be functional over a multitude of topics.

Key to allowing the user to update the word strength list, SentiStrength has developed an algorithm to fine-tune the sentiment strengths using a set of training data. This training algorithm to optimize the sentiment word strengths uses baseline human-allocated term strengths to assess whether an increase or decrease in the term strength of the word strength list would increase the accuracy of the classifications. Table 1 identifies 12 of the 117 words that were optimized. The algorithm tests all words in the word strength list at random and is repeated until all the terms have been verified.

**Table 1. Sample Word List**

| Word | Sentiment Range: -5 to -1, +1 - +5 | | |
|---|---|---|---|
| | Previous Value | Change | Current Value |
| ugh | -3 | +1 | -2 |
| hate | -3 | +1 | -2 |
| heavenly | 4 | -1 | 3 |
| obnoxious | -3 | +1 | -2 |
| petty | -1 | -1 | -2 |
| scary | -2 | -1 | -3 |
| well | 2 | -1 | +1 |
| worse | -3 | +1 | -2 |
| worst | -3 | +1 | -2 |
| wtf | -3 | +1 | -2 |
| yay | 2 | -1 | +1 |

Other term lists are used to help predict the connotation behind a term relative to the context of the text. A booster word list contains words that boost or reduce the emotion of subsequent words, whether positive or negative. A negating word list contains words that invert subsequent emotion words (including any preceding booster words). Negating words include *not, won't, couldn't, shouldn't, wouldn't*, etc. A comprehensive list of emoticons annotated with sentiment values is incorporated in the analysis. The emoticon list assigns values to emoticons; for instance, a smiley sign, :), was assigned a value of 1, while a frown was assigned a value of -1. A questions word list was used to help remove questions that do not contain sentiment when a negative sentiment was indicated. Positive sentiments were not treated in this manner because many question sentences appeared to contain mild positive sentiment (+1) (Thomas et al. 2006).

One of the limitations of SentiStrength is that the sentiment values are highly context-dependent. For instance, sentiment measurements from blog posts may not be appropriate to classify sentiments for social media conversations. Therefore, SentiStrength needed to be calibrated properly before applying it to a specific context.

SentiStrength incorporated algorithms to help correct Tweeting "norms." A spelling correction algorithm was included to correct user tendencies of adding extra letters to a word. For example, "mooooove" would be identified as "move" by the algorithm. If a spelling mistake is made, the algorithm attempts to correct the spelling based on English word list. The algorithm also gives additional strength values of + or -1 when repeated letters are used. Exclamation marks carry a minimum positive strength of 2, and repeated punctuation marks can boost the preceding word or sentence. Few examples of the sentiment calculation of the tweets are given in Figure 1.

Results from the data analysis are presented in several ways. The total positive and negative sentiments were analyzed to identify if a time period had a significant increase or decrease. After identifying significant changes in total positive or negative sentiment over a time period, the specific time period was further scrutinized to identify contributing factors to the change in sentiment. Normalized average sentiment, Equation 1, was computed to identify an average, overall, sentiment. It is important to characterize the overall experience of all riders for a particular time period. The benefit of using a normalized average sentiment is that one or two tweets alone cannot greatly impact the average during a time period. Standard error of the average was calculated to determine the error from the mean during a specific time period, Equation 2.

| I | hate | the | blue | line | Positive | Negative |
|---|---|---|---|---|---|---|
| 0 | -2 | 0 | 0 | 0 | 1 | -2 |

| Power | at | green | line | station | out. | Eerie! | Postive | Negative |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | -2 | 1 | -2 |

| the | blue | line | may | not | be | the | most | glamourous | mode | of | transportation, | but | by | golly | I | don't | envy | the | people | sitting | in | traffic | right | now.. | Positive | Negative |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | -1 |

**Figure 1. Sample tweets and SentiStrength weights**

$$Normalized\ Sentiment\ Strength = \frac{\sum x_i}{n} \quad (1)$$

$$Standard\ Error\ of\ the\ Average = \frac{s}{\sqrt{n}} \quad (2)$$

Where,

$x_i$ = sentiment strength of a tweet

$n$ = total number of tweets

$s$ = standard deviation of the sentiments

Sentiment values for each text were reduced by 1 for positive sentiments and increased by 1 for negative sentiments. This modification to the data is in response to some tweets having both positive and negative sentiments. These tweets tended to be factually based or gave no sentiment. The subtraction or addition of a point of sentiment allows for these tweets to not be included in the total sentiment strength and normalized sentiment strength figures, but to still contribute to the total number of tweets.

Quantifying sentiment is of primary importance, but there is also a need to obtain information from the texts without having to scan each individual text. Tag clouds, or word clouds, are used to better visualize significant words found in the data (Feinburg 2009). The tag clouds are generated by the frequency of words in a body of text. The frequency, or number of occurrences, in which a word appears during a time period can provide details on the overall sentiment or reasons behind a sentiment. Many words are redundant for analysis and can be deleted. For example, in the tag clouds generated in this report, the words *line*, *CTA*, *Chicago*, and *rt* (retweet) were deleted. These words have a high frequency, and their presence may hide more descriptive words.

## Data Collection

A text collection was collected manually for tweets containing the keywords of all combinations of "L" train names near the city of Chicago to optimize, or fine-tune, the terms' positive and negative values in the term strength list. For example, when looking for tweets relative to the Blue Line "L" in Chicago, the search command was given "Blue Line near:Chicago." This method of manual searching was used to collect relative tweets for each of the other seven differently-colored lines. SentiStrength recommends collecting at least 500 texts related to a project in order to use those texts to optimize SentiStrength's text strength list. In this analysis, 557 texts (tweets) were classified and allowed to run through SentiStrength's algorithms to optimize the term strength list.

To observe and collect real time data, the Streaming API method, allowed by Twitter, was used. Streaming API allows members of Twitter to monitor public statuses from all users, filtered in various ways: by user identification number, by keyword, by random sampling, by geographic location, etc. Data were collected using tracking statuses containing the names of each line. After a certain time period, the data were condensed to account for user identification number, location, time, and text. A number of texts were received that contained locations from around the world. Only texts associated with user locations relevant to Chicago and its suburbs were used to ensure that only data relevant to CTA lines were analyzed. There is a chance that a user from Chicago made a comment on another city's transit system but, having reviewed the data, it seems unlikely. Advertisements and GPS monitoring texts, such as FourSquare, were deleted unless a user sentiment was attached to the text. For example, on July 11 at 8:49 AM, an uploaded text "Stuck on the platform :-( (@ CTA - Addison (Red Line)) http://4sq.com/pwzy8y" was not deleted because it contained sentiment in the form of the emoticon. Table 2 lists the dates and time intervals evaluated.

**Table 2. Observed Time Intervals**

| Date | Time Interval | Number of Relevant Tweets |
|---|---|---|
| Monday, July 11, 2011 | 7am–11pm | 158 |
| Tuesday, July 12, 2011 | 8am–11pm | 28 |
| Wednesday, July 13, 2011 | 7am–12pm | 23 |
| Thursday, July 14, 2011 | 11am–12am | 70 |
| Friday, July 15, 2011 | 11am–7pm | 67 |
| Monday, July 04, 2011 | 1pm–9pm | 65 |
| Saturday, July 23, 2011 | 6am–12pm | 46 |

## Sentiment Analysis Results

Figures 2 through 8 illustrate the total positive and negative sentiments, normalized sentiment strength, and total number of tweets. Throughout the graphs of total sentiment, negative sentiment strengths are more abundant. Riders express their sentiment negatively more than positively when an event occurs. The normalized sentiment strength is predominantly negative for each time period because there are higher totals for negative sentiment strength than positive sentiment strength. Unusually high totals for a given hour compared to other hours in the time period give indications of significant incidents.

Data streams from 7/11/2011 contain the most number of tweets in the shortest time period observed, as shown in Table 1. Figure 2 shows that the normalized sentiment strength is continuously negative throughout the interval, yet there remains some positive sentiment strength. The number of incoming tweets reached a maximum at 8:00 and then decreased. The standard error during this time was also the least over the time period. Reasons as to why the maximum occurred at 8:00 are evident in the individual texts. There were numerous tweets containing negative sentiment about delays on the Red, Purple, and Brown lines. The Red line was experiencing power outages while debris, or reportedly a fallen tree, had caused delays on the Purple line. The Brown line appeared to be waiting out the storm before proceeding on its scheduled route. The individual texts indicate the predominant negative sentiments present in the majority of the texts. The individual texts indicate negative sentiment across the performance measures of service and safety. The poor weather had caused delays affecting service while also putting commuters at risk from power outages and debris covered tracks. Figure 9 illustrates the tag cloud generated for 8:00. Words like *red*, *purple*, *brown*, *tree*, *weather*, *fail*, and *yelling* gave details to the trains involved, the reasons for the incident (*weather, tree*), and the general mood (*yelling*).

At 22:00 in Figure 7, there was a strong negative normalized sentiment strength and total negative sentiment. The tweets indicated both a location and a cause of this strongly negative sentiment. This value resulted from fires due to fireworks alongside the Blue line, which caused delays. The tweets indicated that the fires were near the intersection of California St. and Fullerton St. All the information was from tweets collected over this time period. A tag cloud was generated for the 22:00 hour (Figure 10). It too indicates a high frequency for the words *fire, delay,* and *blue*. From the tag cloud, we are able to see problems in relation to security (fire), safety (fire), and mobility (delay). At 20:00, the average sentiment was slightly positive, but only because there was only one tweet recorded at this time period. The lack of a sufficient number of tweets skewed the results by giving the total sentiment for the system based solely on one rider. This fact shows the importance in the number of tweets recorded. A more thorough account of the sentiment is derived from greater participation among the users.

Figure 8 shows a huge spike in total number of tweets at 9:00. The total negative sentiment strength is at a minimum as well at 9:00. After generating a tag cloud (Figure 11), we find words such as *red, blue, flooded, broken,* and *stuck*. The negative sentiments indicate that high rains had caused flooding and delayed service to the Red and Blue lines. There was great dissatisfaction with the lack of service due to this disruption. As early as 6:00, there were tweets coming in that expressed dissatisfaction due to a disruption in the Red line service.

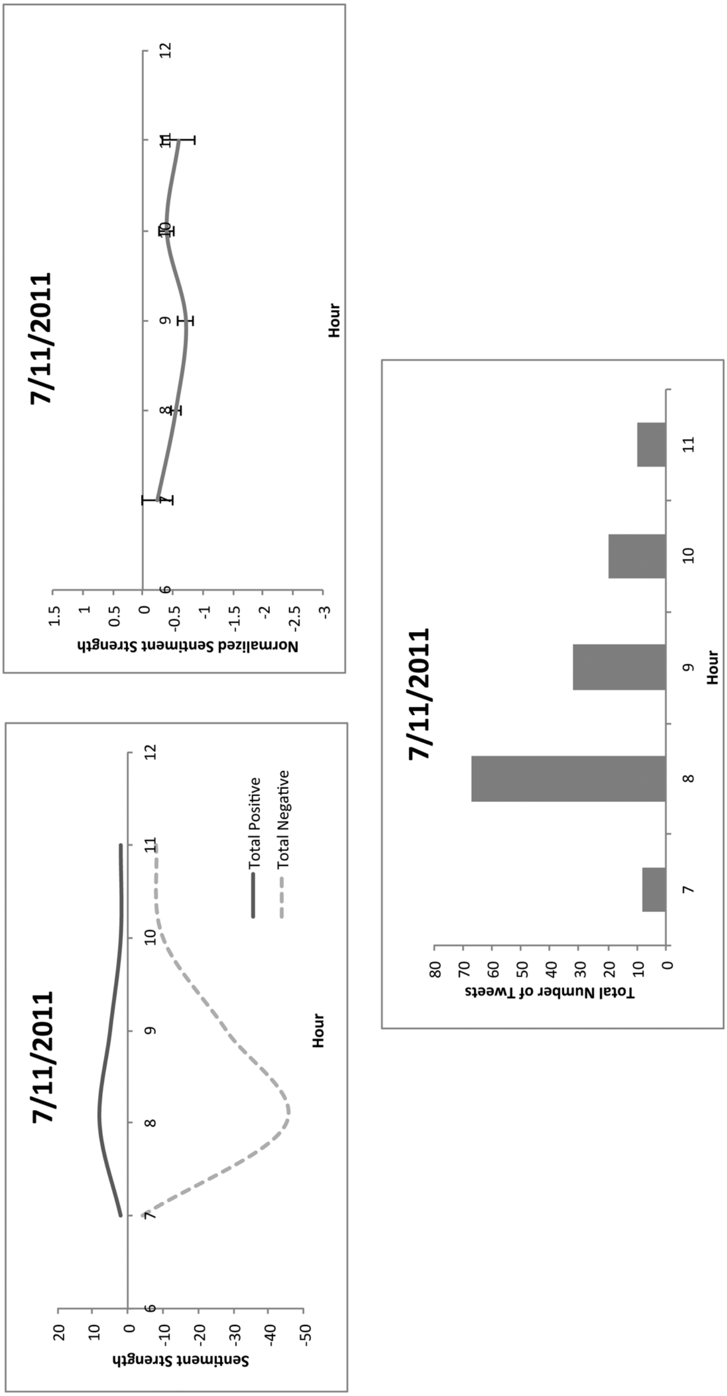



**Figure 2. Total sentiment strengths, normalized sentiment strength, and total number of tweets for 7/11/2011**

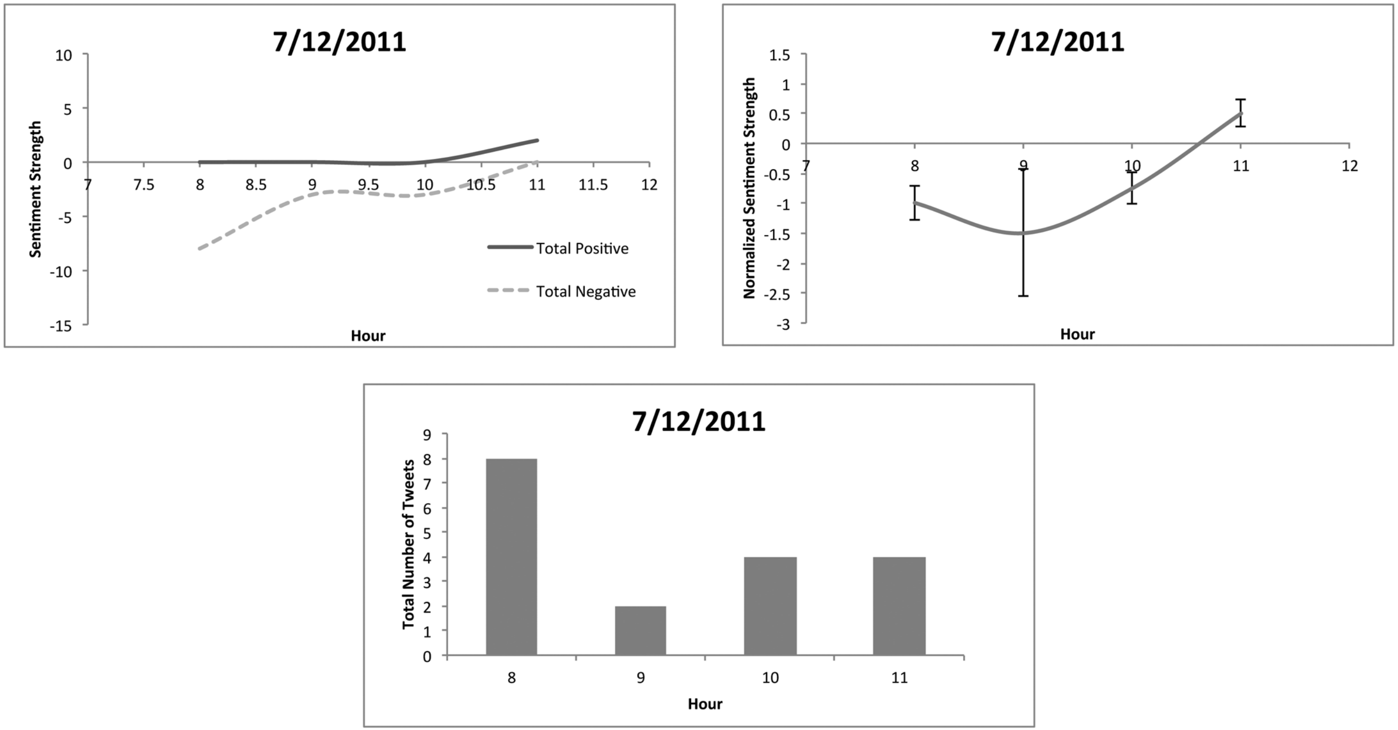



**Figure 3. Total sentiment strengths, normalized sentiment strength, and total number of tweets for 7/12/2011**

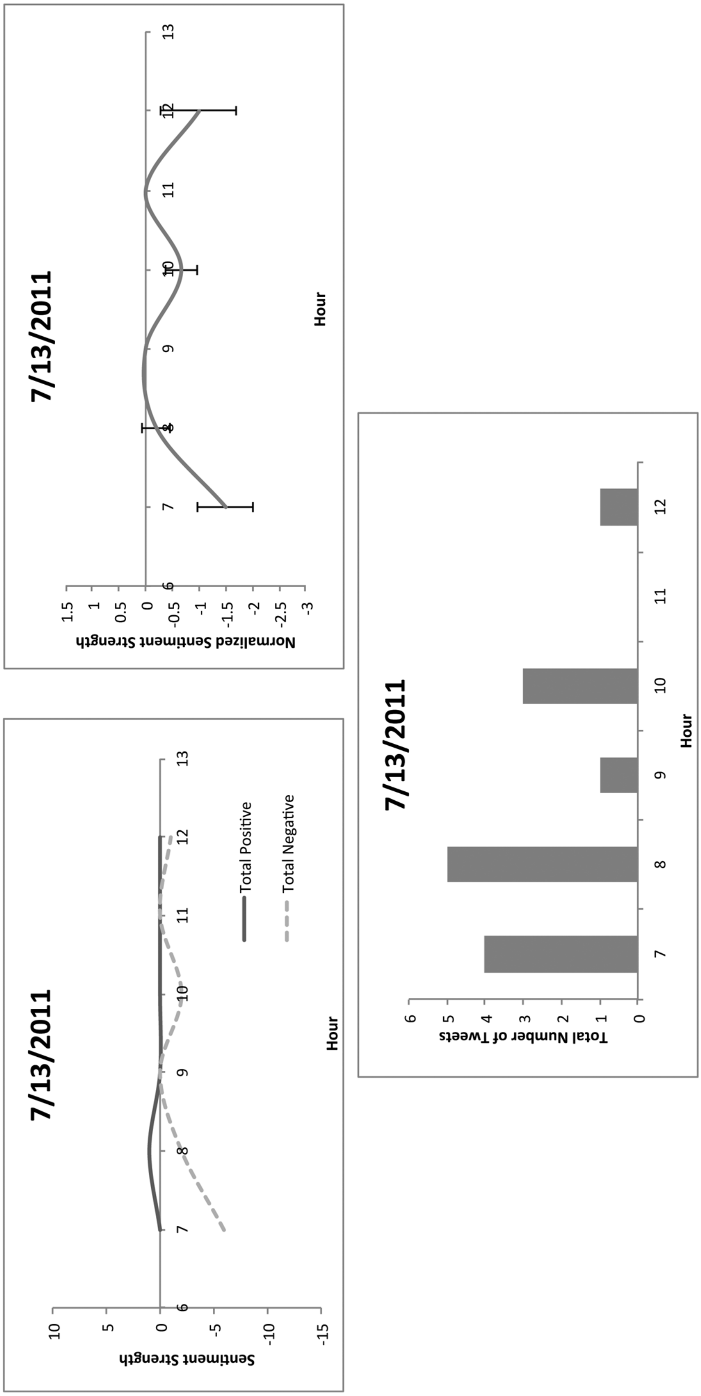



**Figure 4. Total sentiment strengths, normalized sentiment strength, and total number of tweets for 7/13/2011**

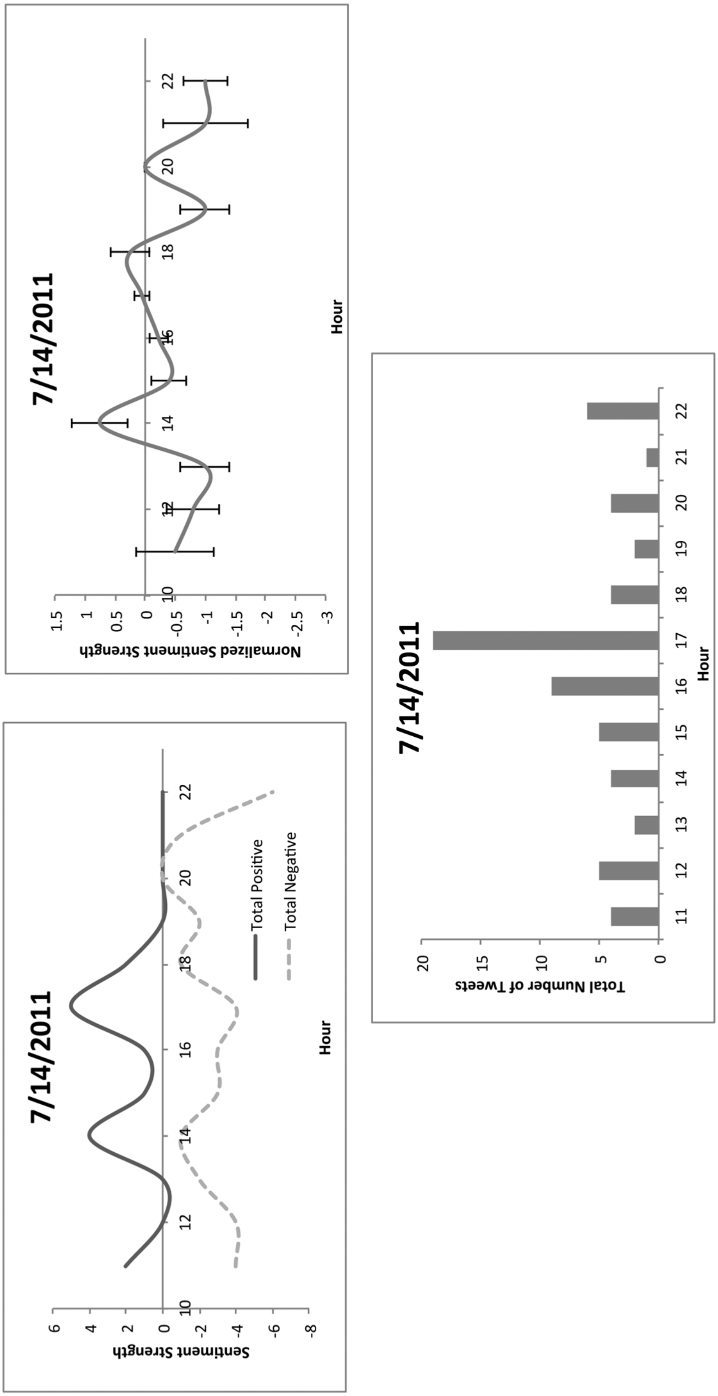



**Figure 5. Total sentiment strengths, normalized sentiment strength, and total number of tweets for 7/14/2011**

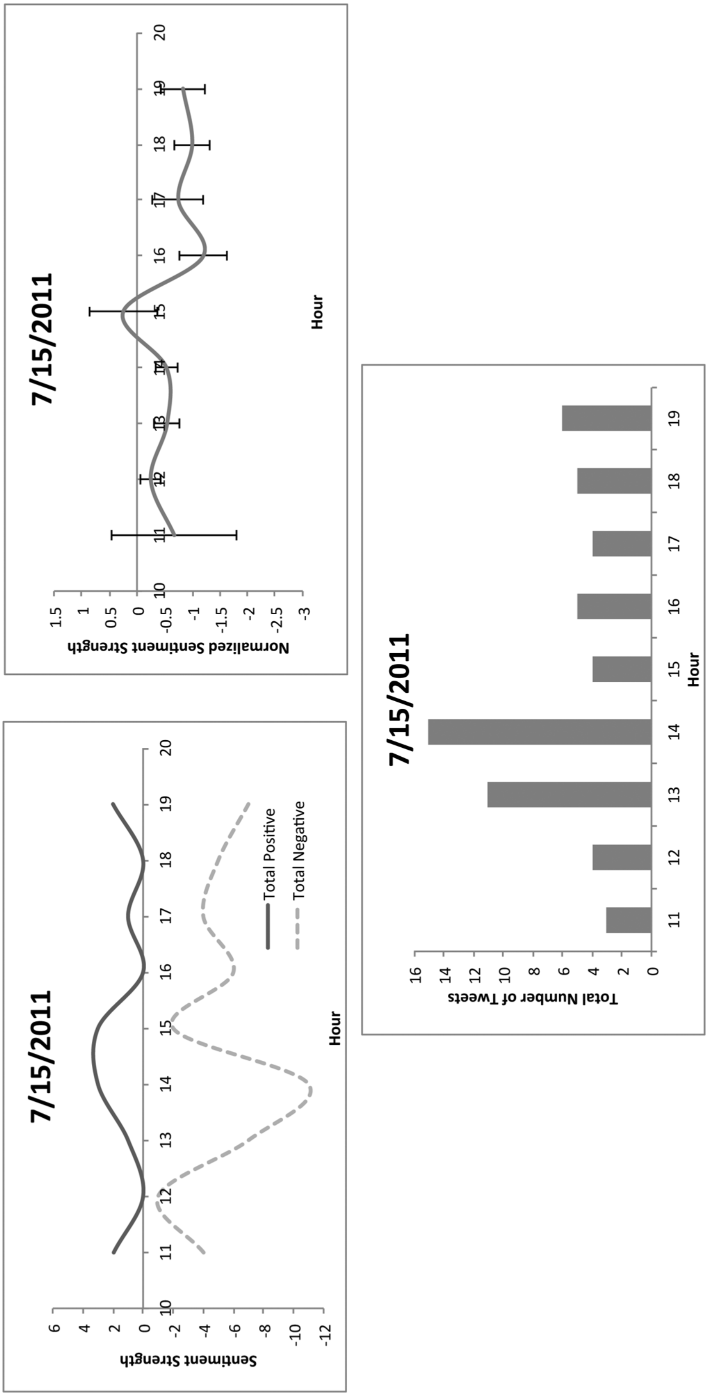
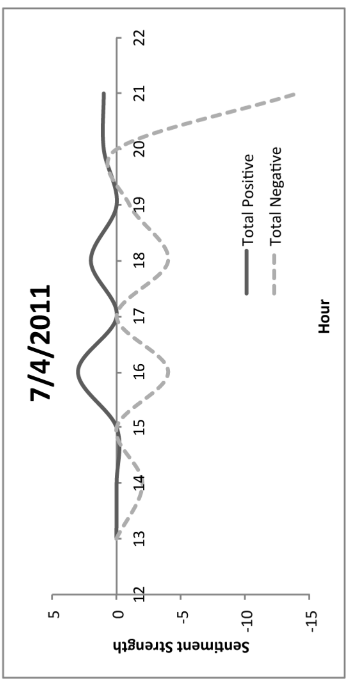
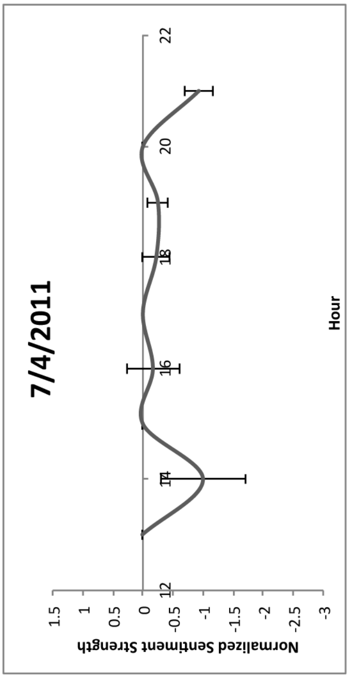



**Figure 6. Total sentiment strengths, normalized sentiment strength, and total number of tweets for 7/15/2011**



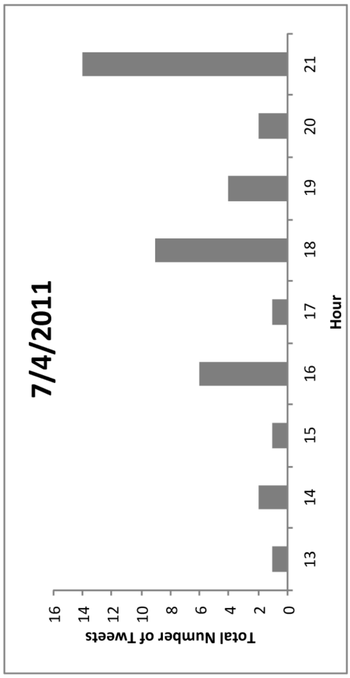




**Figure 7. Total sentiment strengths, normalized sentiment strength, and total number of tweets for 7/04/2011**

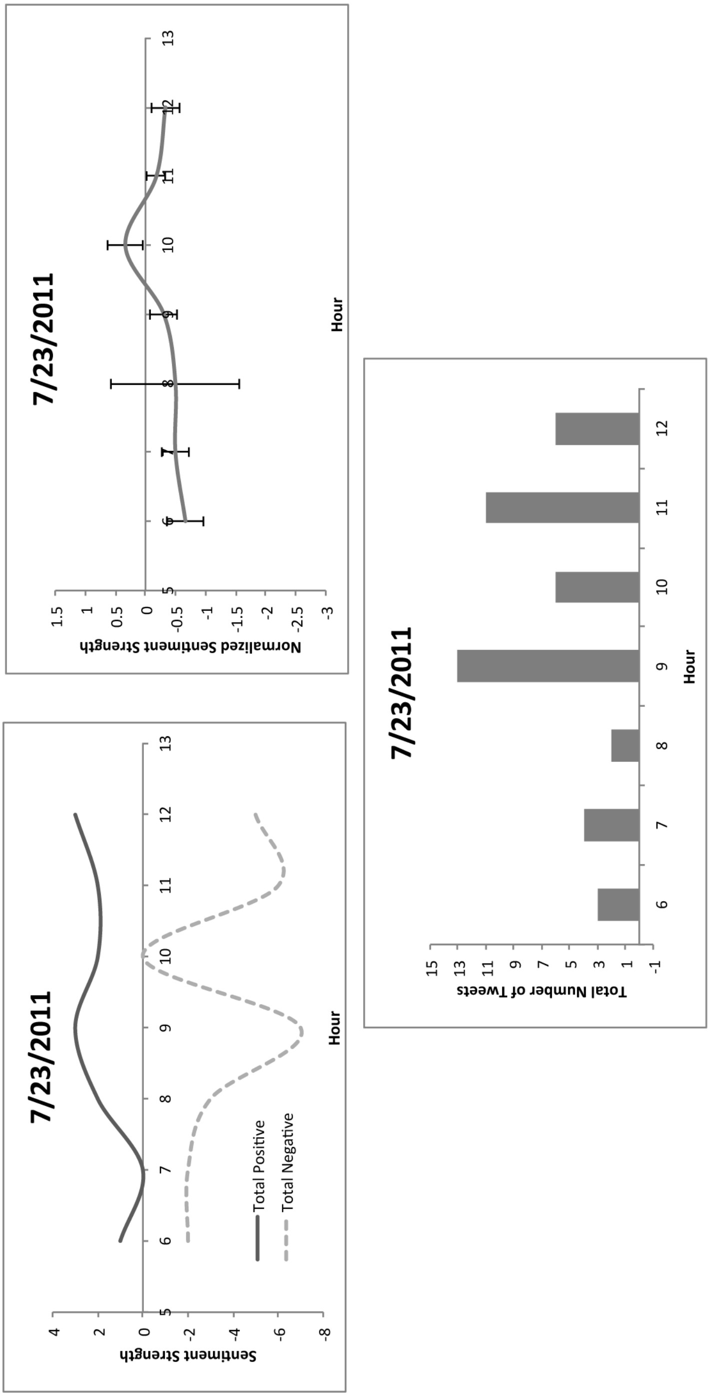



**Figure 8. Total sentiment strengths, normalized sentiment strength, and total number of tweets for 7/23/2011**


Figure 9. Tag cloud for 7/11/2011


Figure 10. Tag cloud for 7/04/2011


Figure 11. Tag cloud for 7/23/2011

## Summary and Conclusions

This paper presents a method for evaluating transit rider satisfaction, which can indirectly indicate the quality of the service provided by the system. The method uses sentiment analysis of transit riders' short messages shared on online social media. The benefits of measuring sentiment in comparison to using traditional surveys for measuring transit riders' satisfactions include:

- The cost of data collection is minimal.
- Data can be collected in real time.
- User-specific needs can be assessed.
- Data can provide meaningful insight as to why a particular sentiment is felt.

Analysis showed that transit riders are more inclined to assert negative sentiments to a situation than positive sentiments. Surely, a lack of total negative sentiment is more desirable to a transit authority than a low total positive sentiment. It is the opinion of the authors that the findings indicate that sentiment analysis can successfully detect rider sentiments in real time of a transit system. Most important, rider dissatisfaction related to such incidents can be quantified through the sentiment analysis of the social media data. Thus, this paper demonstrates a novel method of measuring rider satisfactions using an innovative data source.

This proof of concept should be strengthened with a larger data set to negate possible sample biases in the future. Sample bias should be carefully addressed due to the insufficient amounts of data collected on certain lines, mainly Orange and Pink. The data collected are limited to English only, which may exclude Chicago's large Latino and Polish communities. The city of Chicago provides numerous maps organizing local communities by race, age, English fluency, etc. (City of Chicago 2011). Based on review of a map titled "Distribution of Households Where English is Poorly Spoken," there is a large community of poor English speakers that run alongside both the Orange and Pink lines. Populations of lower income may have been ignored because they do not have access to the technology needed to use Twitter's services due to phone type or limits on available service. It is no secret that Twitter is used more among those below 40 years of age than above. There may also limits on how many older users participated in the present study. A larger sample size would assist in limiting sample bias. Other issues of sample bias are in the form of overlapping tweets from Metra customers. Metra is a rail service that provides transportation outside of the city of Chicago, as well as nationally. There may be certain instances where texts meant for Metra were inferred as being about CTA.

Social media serves as a prediction tool to make informed decisions in every facet of daily life. This research can help transit authorities better operate and manage their system and help riders to have a better travel experience. Obtaining the data is very cost effective and would provide insights not found in traditional performance metrics. The concept of using social media data for measuring perceptions can be applied to other applications where one might want to know the opinions of a large population in real time. Social media can play a role in deciphering the enormous amount of opinions generated daily by filtering the data and then quantifying the opinions on a range of topics. Further research is needed to predict rider destinations and the behavioral factors that influence destination choices. Knowing these factors will serve as a beneficial modeling tool for transit authorities for extending existing networks or services.

An area that needs to be addressed in the future is how social media can be used to enhance a rider's ability to make informed decisions about travel scheduling to travel more efficiently and comfortably. Suppose a rider wants to travel by train, but there are power outages affecting rail lines. Having the ability to receive real-time information on disruptions and disturbances can allow riders to make informed decisions and increase their overall utility.

In the future, exploration into combining opinion mining with this analysis would assist in the manually deletion of advertisements and GPS applications. A continued updating of the word strength list with other available data such as customer hotline data and feedback on transit agency websites and Facebook pages would give valuable addition to the Twitter data.  Developing a tool to assign highly-occurring words over a given time period to a list of specific transit concerns would allow for the describing and quantifying overall sentiment to service, safety, weather, etc., without having to manually investigate individual tweets. Working with a transit authority directly to advertise that sentiment may help increase rider awareness and, thus, increase the total number of tweets received.

## References

Andreassen, T. W. 1995. (Dis)satisfaction with public services: The case of public transportation. *Journal of Services Marketing* 9 (5): 30-41.

Bertini, R. L., and A. El-Geneidy. 2003. Generating transit performance measures with archived data. *Transportation Research Record* 1841: 109-119.

Bollen, J., and H. Mao. 2011. Twitter mood as a stock market predictor. *IEEE Computer* 44(10): 91-94.

Cambridge Systematics. 1999. NCHRP Web Document 26: Multimodal transportation: Development of a performance-based planning process. TRB of the National Academies, Washington, D.C., 1999.

Chicago Transit Authority (CTA). 2011. CTA Monthly Performance Metric. 2011. www.transitchicago.com/assets/1/board_presentations/CTA_Company_Wide_KPI_Mar11.pdf. Accessed May 5, 2011.

City of Chicago. 2011. Census Maps 2011. www.cityofchicago.org/city/en/depts/doit/supp_info/census_maps.html. Accessed November 11, 2011.

Feinburg, J. 2009. Wordle.net  Accessed July 4, 2011.

Finn, A., N. Kushmerick, and B. Smyth. 2002. Genre classification and domain transfer for information filtering. *Proceedings of the European Colloquium on Information Retrieval Research*, Glasgow.

Hanna, J. B., and J. T. Drea. 1998. Understanding and predicting passenger rail travel: An empirical study. *Transportation Journal* 38(1): 38-46.

Hansen, L., A. Arvidsson, F. Nielsen, and E. Colleoni. 2010. Good friends, bad news: Affect and virality in Twitter. *Proceedings of the 2011 International Workshop on Social Computing, Network, and Services*. Technical University of Denmark.

Jindal, N. and B. Liu. 2007. Review spam detection. *Proceedings of the 16th International Conference on World Wide Web,* New York.

Jindal, N. and B. Liu. 2008. Opinion spam and analysis. *Proceedings of the Conference on Web Search and Web Data Mining (WSDM).*

Kim, S. and E. Hovy. 2004. Determining the sentiment of opinions. *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland.

Koushki, P. A., O. I. Al-Saleh, M. Al-Lumaia. 2003. On management's awareness of transit passenger needs. *Transport Policy* 10(1): 17-26.

Metropolitan Transit Authority (MTA). 2011. New York Transit performance dashboard. March 2011. www.mta.info/persdashboard/performance14.html. Accessed June 20, 2011.

Niu, Y., X. Zhu, J. Li., and G. Hirst. 2005. Analysis of polarity information in medical texts. *Proceedings from the American Medical Informatics Association (AMIA) Annual Symposium*, 570-574.

Pang, B. and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1-2): 1-135.

SentiStrength http://sentistrength. wlv.ac.uk. Accessed August 2011.

Thelwall, M., K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology* 61(12): 2544-2558.

Thevathasan, A., and B. Balachandran. 2007. Customers' perceptions of Metropolitan Train Services in Melbourne. 30th Australasian Transport Research Forum (ATRF), Melbourne, Victoria, Australia.

Thomas, M., B. Pang, and L. Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 327-335.

*Transit Capacity and Quality of Service Manual*. 1999. In Transit Cooperative Research Program Research Results Digest, 35: 24-34.

Turney, P. D. 2001. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics*, 417-424.

Zhuang, L., F. Jing, Y. Zhu, and L. Zhang. 2006. Movie review mining and summarization. In *Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM-SIGIR) Conference on Information and Knowledge Management (CIKM).*

## About the Authors

**Craig Collins** (*cbcollin@purdue.edu*) holds bachelor's degrees in Civil Engineering from the School of Civil Engineering and in Industrial Management from the Krannert School of Management at Purdue University. He participated in an internship program in NEXTRANS center in Summer 2011 when he explored a methodology for quantifying transit riders' sentiments through social media.

**Samiul Hasan** (*samiul.hasan@gmail.com*) is a Ph.D. candidate in Transportation Engineering and Infrastructure Systems at Purdue University. He earned master's and bachelor's degrees from the Civil Engineering Department at Bangladesh University of Engineering and Technology in Bangladesh. His research interests include human mobility patterns and activity-travel analysis, machine learning, network modeling, hurricane evacuation and agent-based simulation.

**Satish V. Ukkusuri** (*sukkusur@purdue.edu*) is an Associate Professor in the School of Civil Engineering at Purdue University. He earned a master's degree from University of Illinois Urban-Champaign and a Ph.D. from the University of Texas, Austin. He leads the Interdisciplinary Transportation Modeling and Analytics Lab at Purdue. His current areas of interest include dynamic traffic modeling, network design, use of real-time information for travel time prediction, use of large-scale disaggregate data sets to predict traffic system state, signal control with vehicle to vehicle/infrastructure (V2V and V2I) information exchange, behavioral issues in natural hazards such as hurricanes, evacuation modeling, modeling cap-and-trade policies, modeling policy questions related to plug-in electric vehicles, freight traffic networks, logistics, and safety modeling.