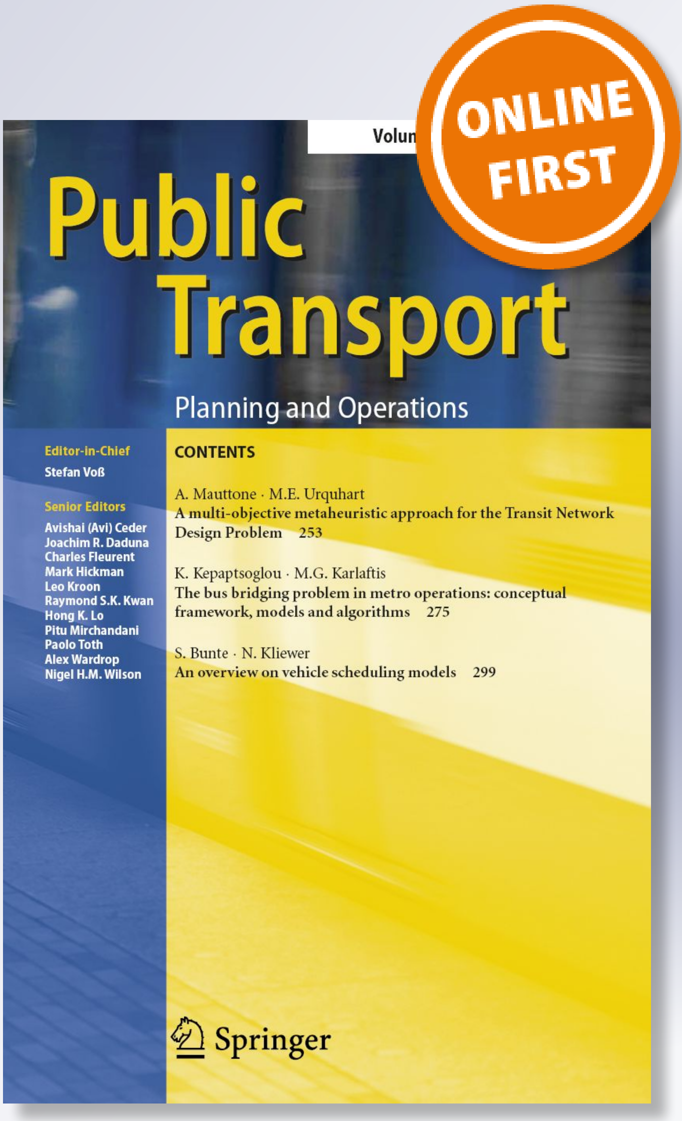
Using Twitter data for transit performance assessment: a framework for evaluating transit riders' opinions about quality of service

N. Nima Haghighi, Xiaoyue Cathy Liu, Ran Wei, Wenwen Li & Hu Shao

Public Transport
Planning and Operations

ISSN 1866-749X

Public Transp
DOI 10.1007/s12469-018-0184-4



Public Transport
https://doi.org/10.1007/s12469-018-0184-4

ORIGINAL PAPER

Using Twitter data for transit performance assessment: a framework for evaluating transit riders' opinions about quality of service

N. Nima Haghighi[1] · Xiaoyue Cathy Liu[1] · Ran Wei[2] · Wenwen Li[3] · Hu Shao[3]

Accepted: 2 August 2018

Abstract
Social media platforms such as Facebook, Instagram, and Twitter have drastically altered the way information is generated and disseminated. These platforms allow their users to report events and express their opinions toward these events. The profusion of data generated through social media has proved to have the potential for improving the efficiency of existing traffic management systems and transportation analytics. This study complements existing literature by proposing a framework to evaluate transit riders' opinion about quality of transit service using Twitter data. Although previous studies used keyword search to extract transit-related tweets, the extracted tweets can still be noisy and might not be relevant to transit quality of service at all. In this study, we leverage topic modeling, an unsupervised machine learning technique, to sift tweets that are relevant to the actual user experience of the transit system. Sentiment analysis is further performed based on the tweet-per-topic index we developed, to gauge transit riders' feedback and explore the underlying reasons causing their dissatisfaction on the service. This framework can be potentially quite useful to transit agencies for user-oriented analysis and to assist with investment decision making.

Keywords Topic modeling · Latent Dirichlet allocation (LDA) · Sentiment analysis · Transit service performance · Quality of transit service

1 Introduction

Social media platforms such as Facebook, Instagram, and Twitter have revolutionized the processes in which information is generated, shared, and stored. With the profusion of Internet of Things (IoT) devices, huge amounts of social media data are created. For instance, Twitter reported that 500 million tweets are sent

✉ Xiaoyue Cathy Liu
cathy.liu@utah.edu

Extended author information available on the last page of the article

Published online: 22 August 2018

each day (Maghrebi et al. 2015). The social media data have drastically altered the way information is disseminated and exchanged (Kaplan and Haenlein 2010). With rich semantic and multimedia content, users of these location-based social media services can be seen as "semantic sensors" with the ability to report and describe events by sending messages with geographic footprints (Goodchild 2007). In other areas, social media data have been used for nowcasting various economic activities. For example, several studies used social media and online searches to nowcast different economic activities (Bughin 2015; Barreira et al. 2013). For example, Bughin (2015) used Google search queries and social media comments to nowcast the product sales evolution of the major telecom companies in Belgium. Barreira et al. (2013) studied the use of Internet search information to improve the nowcasting ability of simple autoregressive models. They used Google Trends to forecast unemployment rates and car sales in four countries including Portugal, Spain, France, and Italy. A social media dataset also presents unprecedented opportunities for creating a cohesive and seamless integration of urban transportation and technology. It has the potential to provide context to transportation performance monitoring and evaluation. Forward-thinking transportation analytics has started to realize the advantages of using such an explosion of data to manage mobility. For example, the city of Los Angeles partnered with Google Waze to extract information from people using this navigation app and learn where congestion hotspots are (Goldsmith 2017). The city also partnered with Esri and developed a geospatial data visualization platform. One of the projects "High Injury Network" maps the city's pedestrian and cyclist fatalities related to traffic incidents to identify risk factors and prevention strategies (Vision Zero 2016). Such developments, integrating the physical transportation assets with virtual structure, allow agencies to improve the traffic management and operations, and the general public to better understand their local environment. More importantly, it will inform evidence-based and data-driven decision-making in transportation policy and investment choices.

Public transit is in direct competition with automobiles. Transit agencies always aim to achieve a compromise regarding the highest ridership possible and the lowest operational costs, as ridership is generally considered as a surrogate measure for revenues (Wei et al. 2017). A myriad of factors can affect transit ridership, including service quality (reliability, comfort, and convenience), service coverage, station accessibility, user experience (Fayyaz et al. 2017; Farber et al. 2016). The current practice for transit agencies to evaluate user experience is to conduct Customer Satisfaction Surveys (CSS) to bus riders. Through these surveys, passengers express their opinions about various attributes describing quality of transit service in terms of a pre-defined scope of evaluation (Transportation Research Board 2003). The high cost, limited sample size, and low resolution have been the major obstacles to make the full use of survey results to inform investment decisions. Moreover, travelers' actual experiences might tell entirely different stories in comparison with these surveys. One alternative to gauge transit riders' experience is through the mining of social media data, to augment the data collected via traditional approaches. Such method is much less costly and time-consuming and allows transit agencies to leverage synergistic benefits for effective transit planning and management.

This study attempts to use Twitter data to evaluate people's opinion about quality of transit service in Salt Lake City, Utah. We developed a framework to effectively extract tweets relevant to public transit service performance, through the use of a machine learning technique—Latent Dirichlet allocation (LDA) model (Blei et al. 2003). A sentiment analysis was then performed to evaluate transit users' feedback on quality of transit service and explore the underlying reasons causing users' dissatisfaction. This framework can be used by transit agencies to evaluate transit service performance from the users' perspective. The results can also help agencies to examine transit-related policy and management.

2 Literature review

A myriad of studies have attempted the use of social media for transportation research. These studies can be classified into four major categories including travel demand estimation (Tasse and Hong 2014; Golder and Macy 2011; Yin et al. 2015), mobility behavior assessment (Cheng et al. 2011; Cho et al. 2011; Hasan et al. 2013), traffic condition monitoring (Tian et al. 2016; Steur 2015; Wanichayapong et al. 2011; Kosala and Adi 2012; Gao et al. 2012), incidents and natural disasters modeling (Sakaki et al. 2010; Lindsay 2011; Ukkusuri et al. 2014; Fu et al. 2015; Mai and Hranac 2013). Only several studies to date have used social media information for public transit analysis, mostly focusing on sentiment analysis to evaluate transit system performance from a transit riders' perspective (Schweitzer 2014; Collins et al. 2013; Luong and Houston 2015).

Schweitzer (2014) used tweets to evaluate users' opinions about public transit. She found that Twitter users express more negative sentiments about public transit than other public services (e.g., police department). Moreover, transit agencies that respond directly to the questions and criticisms of their users demonstrate more positive sentiments. Collins et al. (2013) analyzed Twitter data to assess transit rider's satisfaction using a sentiment strength detection algorithm. They collected tweets containing keywords of train names in the city of Chicago. Their results revealed that transit riders tend to express negative sentiments to a situation (e.g., power outage) rather than positive sentiments. Luong and Houston (2015) conducted sentiment analysis to examine Twitter' users attitudes towards light rail services in Los Angeles. Data were collected using the Search Twitter API around Los Angeles using the names of seven light rail lines. Steiger et al. (2014) used various social media data including Twitter, Foursquare, Instagram, and Flicker to analyze public transit flow and detect major transit hubs in London. They used an LDA model to extract train-related tweets and then applied density-based spatial clustering (DBSCAN) to find clusters with points closely packed together. They found that detected clusters were spatially located along the track segments of London. The results were validated using an overlay of the major rail and public transit network from OpenStreetMap.

These aforementioned studies provided valuable insight into the applications of social media data in public transit analysis. While extracting relevant tweets has significant impacts on the accuracy of results, most of the previous studies only used a simple keyword search to filter transit-related tweets. Yet, based on our preliminary analysis, most of these tweets might not really reflect users' feedback on quality of transit service. Moreover, all studies focused on large cities with a high population where lots of tweets are being posted every day. To the best of our knowledge, no study explored the capacity of social media data for transit analysis in cities of medium or small size.

3 Data

Twitter, as a microblogging platform, allows users to share messages with their followers and also access messages of people that each user follows. A tweet is a message of up to 140 characters (Arias et al. 2013). There are three different types of tweets, including *original tweets*, *replies*, and *retweets*. Original tweets are posted on the sender's profile page, and can be replied or retweeted by other users. Users can also mention other users using @ symbol followed by the specific user name. Words preceded by a # symbol are known as hashtags, and are mostly used to assign a tweet to a specific topic. By clicking on hashtags, users are able to track all the tweets on a specific topic.

The data used in this study were collected from Twitter's real-time streaming API, a resource available to the public to access global streams of tweets. It is worth mentioning that Twitter's real-time streaming only provides 1% of total tweet volume at any given time period. Various filters can be applied to extract tweets within a specific geographic area or containing specific keywords. Both filters must be applied before the 1% sample is being drawn. For the purpose of this study, we first used Salt Lake City as a region filter for tweets retrieval. Then, tweets are filtered to contain various transit-related keywords. After our preliminary analysis on the relevant tweets to determine transit-related keywords, several findings and/or patterns are observed:

(a) People tend to use very specific complaints about specific areas, stations, lines, cars, service, etc. This means that a list of line names/numbers and stop locations (even stop numbers) could be very helpful. For example:

- "blue line headed to SLC for 5:11 SB FR is late! Please hold FR for commuters!"
- "The FrontRunner is really cold & it feels like the AC is on @RideUTA"

(b) People tend to use very location-specific jargon or slang. For example:

- TRAX, SL Central, MAX, WES
- Route numbers (e.g., 703 instead of red line)
- Shortened names: FR = front runner, SJ = South Jordan, PV probably is Pleasant View, 900 S instead of 900 South, the 10 on SE 26th instead of the #10 bus on 26th avenue
- Directions can be important, but the order can vary: NB Blue vs. blue line north

(c) People tend to use unusual language, abbreviations, and symbols on Twitter.

- Shortened words, such as passengers = pssgrs
- Emojis could pose challenges, but did not seem to be very frequent

(d) When people complain or praise a service, they seem to do so directly to that service (@ RideUTA). More general complaints (e.g., about fares, or service quality) seem to be part of semi-public discussions with friends and followers, where the comments seem to be more directed towards political statements (e.g., too expensive, not well funded, poor service). Other times, users will include a general hash tag (e.g., #RideUTA) with their complaints, but it just seems to be like yelling into the void, hoping someone might hear them.

Note that Utah Transit Authority (UTA) is the primary transit provider throughout the Wasatch Front of Utah, in the United States, which includes the metropolitan areas of Salt Lake City, Park City, Provo, Ogden, and Tooele. The agency has a Twitter account named RideUTA. In our study, keywords selected to filter relevant tweets include "@RideUTA", "#UTA", "#UtahTransitAuthority", "#Bus", #TRAX", #BRT, #Rail, #Train, "#FrontRunner", "Blue line", "Red line", "Green line", "S Line", "Orange line", "Yellow line", "FrontRunner", "TRAX".

The filtered dataset comprises 403 tweets from May 23, 2017 to May 31, 2017. Each of these tweets is collected from the Salt Lake City region and contains at least one transit-related keyword. Most of the tweets do not have geo-tags and that is because Twitter users rarely publish locations to their tweets due to privacy concerns. The dataset also includes other information such as timestamp, number of times each tweet has been retweeted or favorited, and user profile (e.g., a users' friends count).

4 Methodology

In this section we present a framework to assess transit performance using transit riders' opinions expressed on Twitter. We argue that although previous studies used keyword search to extract transit-related tweets, the extracted tweets can still be noisy and might not be relevant to transit quality of service. For example, a tweet such as "*@RideUTA votes to SELL the land to @ClearfieldCity (who will sell to Stadler Rail). @fox13 #utpol"* contains a transit-related keyword but it is not relevant to the quality of transit service and in turn is not useful for evaluating user's experience. This study uses an LDA model to sift tweets that are relevant to the actual user experience of the transit system, and can be the most useful for user-oriented analysis and decision-making. The proposed framework consists of three major components. First of all, pre-processing of transit-related tweets is conducted to prepare data for semantic analysis. Then, LDA, a topic modeling algorithm is utilized to extract tweets representing transit riders' opinions about quality of transit service. Lastly, a sentiment analysis is conducted to evaluate transit riders' feedback on transit service. The following sections describe the three components in further detail.

4.1 Pre-processing tweets

Pre-processing is one of the most critical steps in text mining. High dimensionality and large size of textual data are a major obstacle in text mining. In our study, the purpose of pre-processing is to reduce the semantic dimension of tweets. After removing retweets, the semantic content of each tweet is analyzed to remove punctuations and web addresses. In the next step, the remaining contents would go through a natural language processing by applying lower-case conversion, tokenization, and stop word. Lower case conversion converts all words into lower case and consequently reduces the size of unique words. Tokenization divides each tweet into single words. Stop words are words that are not useful for a semantic analysis, for example, "and", "the", and "of". Stop word filtering thus discards useless information to reduce high dimensionality and large size of textual data.

4.2 Topic modeling

In this research, we use an LDA model to detect latent topics in transit-related tweets. LDA is an unsupervised machine learning technique that explores latent topics and associated word groups in a large collection of documents. This method utilizes a "bag of words" model treating each document as a vector of word counts. The idea behind it is that each document is a mixture of topics and each topic is characterized by a probability distribution over a number of words (Blei et al. 2003). Note that tweets are usually short in length and it is difficult to detect topics in such short length. To remedy this, we treat tweets published by the same user as a single document. The following gives an overview of the LDA mathematical basis.

Table 1 provides the notations and parameters used in this study. Figure 1 illustrates the graphical presentation of the LDA model. The LDA model first defines K topics where each topic k is characterized with a φ_k distribution over the collection of words. A document i is then generated by sampling θ_i from a Dirichlet distribution (α) and then choosing each word based on θ_i. To generate each word, LDA first samples a topic $z_{i,j}$ from Multinomial (θ_i), and then selects the word from Multinomial ($\varphi_{z_{i,j}}$). This process is summarized as follows:

Step 1 The topic distribution of the ith document, θ_i is generated from a Dirichlet distribution with parameter α

$$\theta_i \sim \boldsymbol{Dir}(\alpha) \quad for \quad 1 \leq i \leq M \tag{1}$$

Step 2 The word distribution of the kth topic, φ_k is generated from a Dirichlet distribution with parameter β

$$\varphi_k \sim \boldsymbol{Dir}(\beta)\, for\ 1 \leq k \leq K \tag{2}$$

Step 3 The topic of the jth word in the i th document, $z_{i,j}$ is generated from a multinomial distribution with parameter θ_i

$$z_{i,j} \sim \boldsymbol{Multinomial}\left(\theta_i\right) for\ 1 \leq i \leq M\ and\ 1 \leq j \leq N_i \tag{3}$$

Table 1 Descriptions of variables and parameters

Notation	Description
t	Index of tweets
i	Index of documents
k	Index of topics
j	Index of words
α	The prior distribution for topics in a document
β	The prior distribution for words in a topic
θ_i	Topic distribution of document i
φ_k	Word distribution of topic k
$w_{i,j}$	j th word in document i
$z_{i,j}$	Topic of word $w_{i,j}$ from document i
K	Number of topics
T	Number of tweets
M	Number of documents
N_t	Number of words in tweet t
N_i	Number of words in document i
P_{tk}	Portion of topic k in tweet t
I_{ik}	Tweet-per-topic k index of tweet t

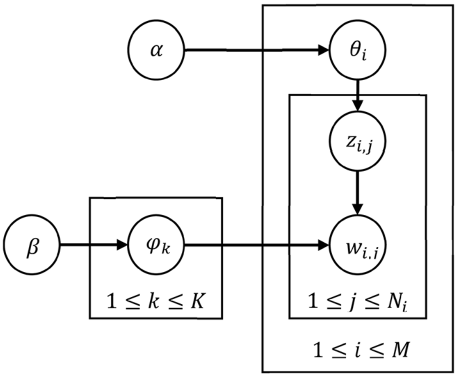


Fig. 1 Graphical representation of the LDA model

Step 4 The j th word in the i th document, $w_{i,j}$ is generated from a multinomial distribution with parameter $\varphi_{z_{i,j}}$

$$w_{i,j} \sim \boldsymbol{Multinomial}\left(\varphi_{z_{i,j}}\right) \, for \; 1 \le i \le M \; and \; 1 \le j \le N_i \tag{4}$$

According to the data generating process, the joint probability of the model is

$$p(w, z, \theta, \varphi | \alpha, \beta) = p(w|\varphi, z)p(\varphi|\beta)p(z|\theta)p(\theta|\alpha) \tag{5}$$

The model training process is to find posterior of document-topic distribution θ and topic-word distribution φ that maximize the model joint probability. The LDA model's approximate inference can be computed by Markov Chain Monte Carlo methods or Variational Expectation Maximization (VEM).

After LDA model estimates the document-topic and topic-word distributions, tweet-topic distribution is then calculated based on the topic-word distribution. Given that each tweet is made up of several words, tweet-per-topic probabilities can be estimated based on the sum of its word probabilities. Equation (6) shows the probability that the t^{th} tweet being generated from topic k.

$$P_{t,k} = \frac{1}{N_t}\sum_{j=1}^{N_t} z_{t,j}\ for\ 1 \le t \le T\ and\ 1 \le k \le K \tag{6}$$

Tweet-topic distribution shows the probabilities that each tweet is being generated from different topics. In order to assign a specific topic to each tweet, we define a tweet-per-topic index $I_{t,k}$. $I_{t,k}$ compares the proportion of topic k with the proportion of other latent topics in the t^{th} tweet. Tweet t with $I_{t,k}$ greater than a specific threshold will be assigned to topic k. Tweet-per-topic index is calculated as follows:

$$I_{t,k} = \log_2 \frac{P_{t,k}}{\left(\sum_{k=1}^{K} P_{t,k}\right) - P_{t,k}} \tag{7}$$

4.3 Sentiment analysis

Sentiment analysis or opinion mining refers to the technique that analyzes people's opinions, sentiments, assessments, attitudes, and emotions towards products or services (Liu 2012). Common methods to conduct sentiment analysis include machine learning based and lexicon-based. The machine learning approach treats sentiment analysis as a text classification problem. These algorithms use a training dataset to learn the model behavior and then make predictions on the testing dataset. The lexicon-based approach estimates the sentiment of each document by scoring each word based on a collection of positive and negative words.

In this study, "Rsentiment" package in R is used for sentiment analysis. The "Rsentiment" package calculates the score of each tweet based on the presence of positive and negative words, presence of negation and sarcasm. The positive and negative scores indicate positive and negative sentiments, respectively. Scores of 0 and 99 represents neutral and sarcasm sentiments, respectively. For further information on Rsentiment package, please refer to Bose et al. (2017).

5 Results

"topicmodels" package in R (Hornik and Grün 2011) is used to run the LDA model. VEM algorithm (Blei et al. 2003) is used to train the LDA model with a different number of topics. The VEM is an efficient algorithm to estimate marginal likelihood of probabilistic models with latent variables (z and θ in the LDA model) or incomplete data. This algorithm defines a lower bound on the marginal likelihood using variational calculus to determine maximum likelihood of the LDA model. For further information on VEM see Blei et al. (2003) and Bernardo et al. (2003). To fit the LDA model, the number of topics needs to be specified in advance. In order to identify the model with the best fit, the log likelihoods of LDA models with a different number of topics are calculated. Results reveal that increasing the number of topics beyond two does not significantly change the model's log-likelihood. Consequently, two distinct topics are detected across all tweets we collected. Figure 2 illustrates the top 15 words that are most commonly used within each topic. The horizontal axis shows the probability that each word is being generated from each topic.

Examining the words within each topic, it shows that topic 1 includes a set of words such as "train", "bus", and "trax", which refer to various transit infrastructures within the Salt Lake region along with words like "late", "schedule", "stop". Therefore, topic 1 seems to include tweets that are relevant to transit service performance. Topic 2 includes words such as "#utpol", "@fox13", "land", "project", "sale", "deal", and "board". Note that #utpol is a hashtag commonly used for sharing Utah political news on Twitter and "@fox13" is the official Twitter account of KSTU-TV in Salt Lake City, Utah that reports Utah local news. Consequently, topic 2 represents tweets discussing news related to UTA projects and decisions.

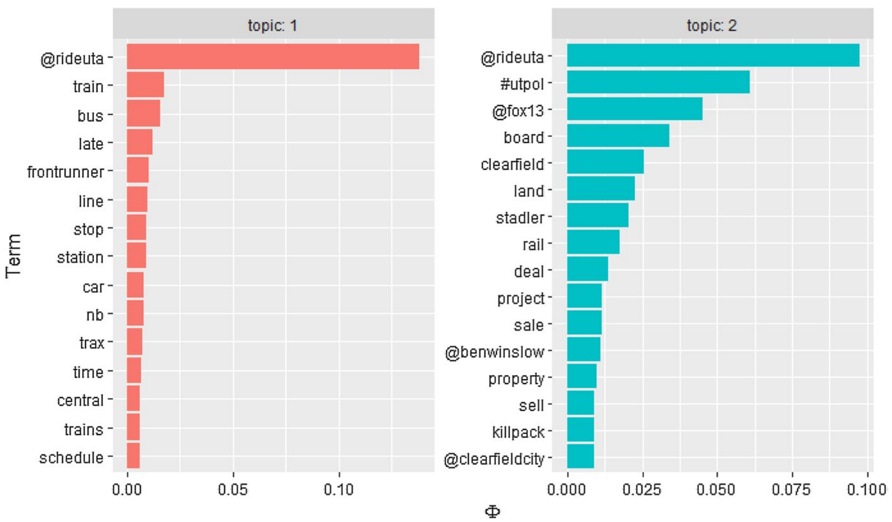


Fig. 2 Frequently used words in each topic based on LDA

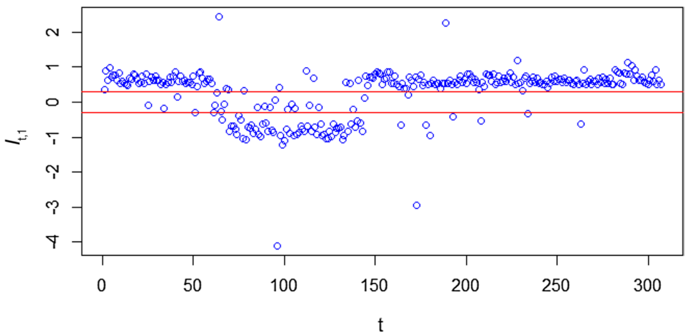


Fig. 3 Tweet-per-topic index of different tweets

Table 2 Sample examples of tweets in each topic based on LDA model

Topic	Tweet
1	@RideUTA Can you have frontrunner wait for your late blue line train at SB Murray central? The train broke down for… https://t.co/5bSJ4cgyH7
1	@RideUTA instead of Frontrunner stopping a mile south to wait for the north train wait for the late trax @ the SLC central station #winning
1	@RideUTA the crossing arms are stuck in the down position at the Layton station crossing
2	#TBT to our project with the @utahdot to build a replacement bridge utilizing Mirafi® geosynthetic reinforcement. https://t.co/GytSF8bVEa
2	@RideUTA votes to SELL the land to @ClearfieldCity (who will sell to Stadler Rail). @fox13 #utpol
2	Stadler Rail letter to @RideUTA says former board member, Sheldon Killpack, not benefit from Clearfield chosen over other sites #utpol

In order to assign a specific topic to each tweet, we calculate the tweet-per-topic index. As explained in Eq. (7), $I_{t,1}$ can be used to extract those tweets expressing transit riders' opinions about quality of transit service (topic 1). If the index $I_{t,1}$ value is around zero, it indicates that the proportions of topic 1 and topic 2 in tweet t are almost identical and we are not able to assign a distinct topic to that tweet. Figure 3 shows $I_{t,1}$ for different tweets. Boundaries are plotted to filter out tweets that can be assigned to neither topic 1 nor topic 2. For the tweets that fall within the boundary region, there is no significant difference between proportions of topic 1 and topic 2. These tweets are discarded from the analysis. The rest of tweets with positive $I_{t,1}$ are assigned to topic 1, and are used for sentiment analysis.

Table 2 shows sample examples of the tweets assigned to each topic. It is clear that tweets assigned to topic 1 are representing the actual user experience of the transit system and can be potentially useful for analyzing users' feedback on quality of transit service. On the other hand, tweets assigned to topic 2 are mostly discussing about UTA decisions and projects.

Figure 4 illustrates the temporal distribution of tweets assigned to topic 1. Note that the peak periods for tweets frequency do not coincide with transit service

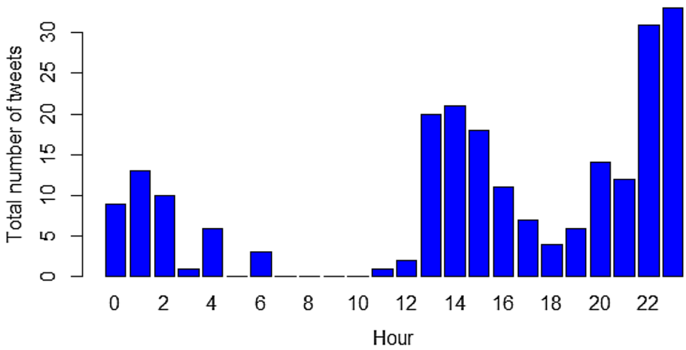


Fig. 4 Temporal distribution of tweets assigned to topic 1

peak periods. However, there is a huge spike in total number of tweets during late night where people might have more free time to talk about their experience during the day.

The sentiment of each tweet is calculated using the "Rsentiment" package. Based on the assigned scores, tweets are classified into 3 categories including *negative*, *positive*, and *neutral*. No tweet with sarcasm sentiment is found. Figure 5 illustrates the number of tweets within each category during each day. As shown in the figure, more transit-related tweets are posted in the middle of the week. Although a largest number of negative tweets were posted on Tuesday, May 23, the proportion of daily negative tweets is greater during weekend than weekdays (e.g., 50% on Sunday, May 28, vs. 36% on Tuesday, May 23). That might be explained by the less frequent service during weekends than weekdays. Examining tweets posted on Tuesday, May 23 revealed that most of the negative tweets were related to the *Frontrunner* (i.e., commuter rail operated by the UTA).

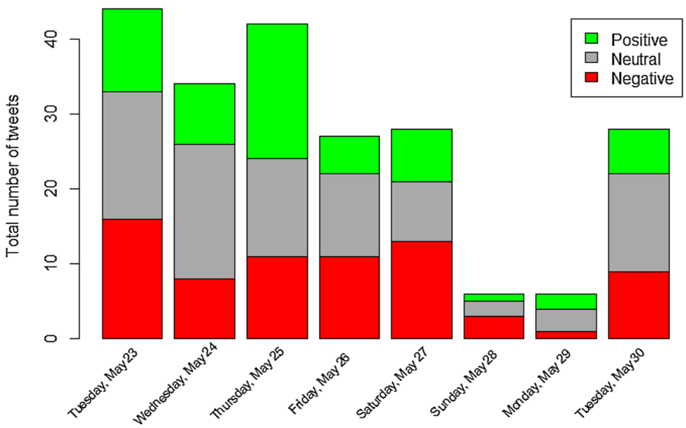


Fig. 5 Histogram of positive, negative and neutral tweets on each day

Fig. 6 Word cloud of tweets with negative sentiments

Table 3 Frequently occurred terms in tweets with negative sentiments

Ranking	Term	Occurrence frequency
1	Train, trains	17
2	Stop	13
3	Late	11
4	Bus	10
5	Station	9
6	Frontrunner, front runner	9
7	Central	6
8	Trax	5

Tweets with negative sentiments are further extracted to explore the underlying reasons that cause riders to express their opinions negatively. Figure 6 shows the word cloud of tweets with negative sentiments. The size of each word indicates the frequency of that word. In order to better compare various word frequencies, the term *@RideUTA* is removed from word cloud. Table 3 summarizes terms with high occurrence frequency in tweets with negative sentiments.

According to Fig. 6 and Table 2, negative sentiments are mostly related to the performance of train and bus systems. It is clear that there are several negative tweets about *FrontRunner* and *TRAX*. *FrontRunner* is a commuter rail operated by UTA serving more than 16,000 riders each day. There are 16 stations along the *FrontRunner* line. The North Temple, Salt Lake Central, and Murray Central stations connect *FrontRunner* to *TRAX* light rail system. *Orem central station*, *Murray central station*, *Draper station*, and *Ogden station* are *FrontRunner* stations which appeared in the word cloud. *TRAX* is a light rail transit system operated by UTA serving Salt Lake County. *TRAX* has three lines (blue, green, and red) and 50 stations with more than 63,000 daily ridership. The word "Blue" in the word cloud is mostly likely referring to *TRAX* blue line.

Blue line goes through Salt Lake City downtown and is among the routes with the highest ridership in the State of Utah.

6 Conclusion

In this study, we propose a framework to evaluate transit rider opinions about quality of transit service. We use text mining techniques to analyze tweets posted within the Salt Lake region between May 23, 2017 and May 31, 2017. The combination of LDA and sentiment analysis enables the tweets on people's opinion about quality of transit service to be extracted and evaluated. Using our sampled tweets, the LDA model separates them into two distinct topics: transit service performance vs. UTA projects or news discussion. The tweets of the first topic might be potentially useful for user-oriented analysis and to assist with investment decision-making. Results of sentiment analysis reveal that the percentage of negative tweets are greater during weekends than weekdays. That might be due to the less frequent service during weekends. Moreover, most of the negative tweets are related to transit routes with high ridership.

Our findings verify the potential of social media data in analyzing quality of transit services. Yet various sources of sampling biases persist that need to be addressed in future studies. For instance, we only focused on tweets in English which might have excluded members of other communities. Also, low income people and senior residents are less likely to use smartphones, which represent a significant portion of the missing inputs in our analysis. Future studies might consider to combine the Twitter analytics with traditional data sources (e.g., census) to explore sample representativeness and interpolation methods to effectively integrate these heterogeneous data sources. This study is exploratory in nature, and collecting tweets over a longer time period (e.g., 1 year) might help address the sample size limitation and remedy the sample bias.

In the future, one can expand on a computational module to assign frequently used terms to a list of specific issues, as a second tier classification based on various attributes such as weather, infrastructures, etc., without manually examining each tweet. Another area that needs to be addressed is how the users' experience extracted from social media can be used in conjunction with transit operational/connectivity analysis to facilitate knowledge discovery. Specifically, users' experience and transit system performance (e.g., delay, accessibility) can potentially complement each other to present a holistic and comprehensive picture to the agencies about how the existing system is functioning. This can be achieved through cross validation and cross-referencing the results generated from the two fronts. Moreover, further research on predicting transit riders' destinations via social media may assist transit authorities to effectively expand existing networks via latent demand estimation.

References

Arias M, Arratia A, Xuriguera R (2013) Forecasting with Twitter data. ACM Trans Intell Syst Technol 5(1):1–24

Barreira N, Godinho P, Melo P (2013) Nowcasting unemployment rate and new car sales in South-western Europe with Google Trends. NETNOMICS Econ Res Electron Netw 14(3):129–165
Bernardo JM, Bayarri MJ, Berger JO, Dawid AP, Heckerman D, Smith AFM, West M (2003) The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. Bayesian Stat 7:453–464
Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
Bose S, Saha U, Kar D, Goswami S, Nayak AK, Chakrabarti S (2017) RSentiment: a tool to extract meaningful insights from textual reviews. In: Proceedings of the 5th international conference on frontiers in intelligent computing: theory and applications, Singapore
Bughin J (2015) Google searches and Twitter mood: nowcasting telecom sales performance. NETNOMICS Econ Res Electron Netw 16(1–2):87–105
Cheng Z, Caverlee J, Lee KD, Sui DZ (2011) Exploring millions of footprints in location sharing services. International AAAI conference on web and social media (ICWSM), pp 81-88
Cho E, Myers SA, Leskovec J (2011) Friendship and mobility: user movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. ACM: 1082–1090
Collins C, Hasan S, Ukkusuri SV (2013) A novel transit rider satisfaction metric. J Public Transp 16(2):21–45
Farber S, Ritter B, Fu L (2016) Space–time mismatch between transit service and observed travel patterns in the Wasatch Front, Utah: a social equity perspective. Travel Behav Soc 4:40–48
Fayyaz SK, Liu XC, Porter RJ (2017) Dynamic transit accessibility and transit gap causality analysis. J Transp Geogr 59:27–39
Fu K, Nune R, Tao JX (2015) Social media data analysis for traffic incident detection and management. Transportation research board 94th annual meeting 15-4022, Washington, D.C
Gao H, Tang J, Liu H (2012) Exploring social-historical ties on location-based social networks. In: International AAAI conference on web and social media (ICWSM). The AAAI Press, California
Golder SA, Macy MW (2011) Diurnal and seasonal mood vary with work, sleep and daylength across diverse cultures. Science 333(6051):1878–1881
Goldsmith S (2017) L.A.'s testing ground for transportation efficiency, Mar. 2016. http://www.governing.com/blogs/bfc/gov-los-angeles-transportation-efficiency-mobility-management.html. Accessed 20 Jul 2017
Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. GeoJournal 69(4):211–221
Hasan S, Zhan X, Ukkusuri SV (2013) Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In: Proceedings of the 2nd ACM international workshop on urban computing, pp 6:1–6:8
Hornik K, Grün B (2011) Topicmodels: an R package for fitting topic models. J Stat Softw 40(13):1–30
Kaplan AM, Haenlein M (2010) Users of the world, unite! The challenges and opportunities of Social Media. Bus Horiz 53(1):59–68
Kosala R, Adi E (2012) Harvesting real time traffic information from Twitter. Procedia Eng 50:1–11
Lindsay BR (2011) Social media and disasters: current uses, future options, and policy considerations. Congress research service 41987
Liu B (2012) Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, vol 5, no 1. Morgan & Claypool Publishers
Luong TT, Houston D (2015) Public opinions of light rail service in Los Angeles, an analysis using Twitter Data. iConference 2015 Proceedings, Philadelphia
Maghrebi M, Abbasi A, Rashidi TH, Waller ST (2015) Complementing travel diary surveys with Twitter data: application of text mining techniques on activity location, type and time. 18th international conference on intelligent transportation systems (ITSC), Las Palmas, Spain
Mai E, Hranac R (2013) Twitter interactions as a data source for transportation incidents. Presented at Transportation Research Board 92nd annual meeting, Washington, D.C
Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World Wide Web, Raleigh, North Carolina
Schweitzer L (2014) Planning and social media: a case study of public transit and stigma on Twitter. J Am Plan Assoc 80(3):218–238
Steiger E, Ellersiek T, Zipf A (2014) Explorative public transport flow analysis from uncertain social media data. In: Proceedings of the 3rd ACM SIGSPATIAL international workshop on crowd

sourced and volunteered geographic information—GeoCrowd'14. New York, New York, ACM Press, pp 1–7
Steur RJ (2015) Twitter as a spatio-temporal source for incident management. Master's Thesis, Utrecht University, Netherlands
Tasse D, Hong JI (2014) Using social media data to understand cities. In: Proceedings of NSF workshop on big data and urban informatics. Carnegie Mellon University, Pittsburg, Pennsylvania
Tian Y, Zmud M, Chiu YC, Carey D, Dale J, Smarda D, Lehr R, James R (2016) Quality assessment of social media traffic reports—a field study in Austin, Texas. Transportation Research Board 95th annual meeting, No. 16-6852, Washington, D.C
Transportation Research Board (2003) Transit capacity and quality of service manual. TCRP Report 100. National Academy Press, Washington, D.C.
Ukkusuri S, Zhan X, Sadri A, Ye Q (2014) Use of social media data to explore crisis informatics: study of 2013 Oklahoma Tornado. Transp Res Rec J Transp Res Board 2459:110–118
Vision Zero (2016) High injury network. http://visionzero.lacity.org/high-injury-network/. Accessed 20 Jul 2017
Wanichayapong N, Pruthipunyaskul W, Pattara-Atikom W, Chaovalit P (2011) Social-based traffic information extraction and classification. 2011 IEEE 11th international conference on ITS telecommunications, pp 107–112
Wei R, Liu X, Wang L, Golub A, Farber S (2017) Evaluating public transit services for operational efficiency and access equity. J Transp Geogr 65:70–79
Yin Z, Fabbri D, Rosenbloom ST, Malin B (2015) A scalable framework to detect personal health mentions on Twitter. J Med Internet Res 17(6):e138

Affiliations

N. Nima Haghighi[1] · Xiaoyue Cathy Liu[1] · Ran Wei[2] · Wenwen Li[3] · Hu Shao[3]

N. Nima Haghighi
nima.haghighi@utah.edu

Ran Wei
ran.wei@ucr.edu

Wenwen Li
wenwen@asu.edu

[1] Department of Civil and Environmental Engineering, University of Utah, 110 Central Campus Drive, Suite 2000B, Salt Lake City, UT 84112, USA

[2] School of Public Policy, University of California at Riverside, 900 University Ave, Riverside, CA 92521, USA

[3] School of Geographical Sciences and Urban Planning, Arizona State University, 975 S. Myrtle Ave, Tempe, AZ 85287, USA