IATSS Research xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

IATSS Research

Research Article

Extracting patterns from Twitter to promote biking

Subasish Das[a,*], Anandi Dutta[b], Gabriella Medina[c], Lisa Minjares-Kyle[d], Zachary Elgart[e]

[a] Texas A&M Transportation Institute, 3135 TAMU, College Station, TX 77843-3135, United States
[b] Computer Science and Engineering, Texas A&M University, College Station, TX 77843, United States
[c] Texas A&M Transportation Institute, Texas A&M University System, 505 E. Huntland Dr. Suite 455, Austin, TX 78752, United States
[d] Texas A&M Transportation Institute, Texas A&M University System, 701 N. Post Oak Rd. Suite 430, Houston, TX 77024, United States
[e] Transit Mobility Program, Texas A&M Transportation Institute, Texas A&M University System, 701 North Post Oak Road, Suite 430, Houston, TX 77024, United States

ARTICLE INFO

Article history:
Received 17 December 2017
Received in revised form 5 August 2018
Accepted 28 September 2018
Available online xxxx

Keywords:
Bike commuting
Text mining
Sentiment analysis
Polarity
Twitter
Social media

ABSTRACT

Emphasis on non-motorized travel modes (for example, biking) reduces motorized trips and provides positive effects on the environment and the quality of human life. Understanding factors that influence people to biking or bike commuting can help decision makers, transportation planners, and bike commuting networks. Historically, conventional methods like surveys and crash data analyses were conducted to understand relevant factors. Survey and crash data analysis are difficult to perform in broad scale due to data availability and efforts. An innovative approach to determining these factors is to conduct social media mining to understand sentiments or motivations of bike commuters. People use terms (with hashtag at the beginning of the term) in Twitter, a popular social media network, to express their thoughts, activities or information. This study developed a framework for using Twitter data in understating the sentiments of the bikers with minimal effort. In this study, Twitter data associated with bike commuting hashtags were obtained for eight years (2009–2016). This study provided a framework of data collection and application of various natural language processing (NLP) tools (for example, text mining, sentiment analysis) to extract knowledge from the unstructured text data. Findings show that biking is associated with weather and seasonal patterns. The general sentiment towards biking is positive. However, negative sentiments are associated with bad weather, crime, and other challenges. The polarity scores indicate somewhat positiveness in the recent few years. The developed framework and the findings of this study will help planners and decision makers to promote biking on a broader scale.

1. Introduction

According to the American Community Survey (ACS), the number of U.S. bicycle commuters increased from about 488,000 in 2000 to about 786,000 (around 60% increase) in 2008–2012, the largest percentage increase than that of any other commuting mode [1]. Although many cities have been investing in bicycle friendly roadway environments in the recent years, limited data availability makes the quantification of mode-choice metrics difficult. The ACS provides one of the nation's most robust bike commuting data but does not capture motivations of bike commuters. Social media mining may present an opportunity to mitigate the lack of resources to understand the motivation and trend towards bike commuting that is not included in ACS data. Twitter is a popular microblogging social media ecosystem used to initiate discussion on a variety of interests. Understanding interactions among users could be a useful tool to understand both disruptive and beneficial topics and trends for policy makers and experts in different domains. This study aims to investigate bike commuting related tweets to extract valuable knowledge from unstructured text data. To accomplish the research goals, Twitter data associated with bike commuting hashtags were obtained for eight years (2009–2016). The final dataset contains around 80,000 tweets. This study employed different natural language processing (NLP) tools to perform the analysis. This paper fills a gap in transportation research by presenting a concept of using social media analytic tools to provide insight into microblogging conversations, such as Twitter. This study has two unique contributions: 1) it developed a framework for collecting and analyzing biking associated texts from Twitter, and 2) it performed a comprehensive sentiment analysis using approximately 10 million words.

2. Literature review

While a decent body of literature on bicycle commuters exists, there is a sparse amount of literature on social media data mining by transportation professionals and even less where the two overlaps. The

[*] Corresponding author.
E-mail addresses: s-das@tti.tamu.edu (S. Das), g-medina@tti.tamu.edu (G. Medina), l-minjares@tti.tamu.edu (L. Minjares-Kyle), z-elgart@tti.tamu.edu (Z. Elgart).
Peer review under responsibility of International Association of Traffic and Safety Sciences.

https://doi.org/10.1016/j.iatssr.2018.09.002
0386-1112/

Please cite this article as: S. Das, et al., Extracting patterns from Twitter to promote biking, IATSS Research (2018), https://doi.org/10.1016/j.iatssr.2018.09.002

Table 1
Studies on transportation related social media content analysis.

No.	Country	Study Period	Research Hypothesis	Findings	Method	Ref
1	Spain	2015	Electronic Word of Mouth (e-WOM)	Influencers use more hashtags and mentions; they tend to have less links in their posts; share opinions and feelings-positive or negative	Detect network threads	5
2	U.S.	2010–2011	How can social networks and social media be leveraged to engage broader participation?	Micro-participation could be effective in generating participation, but there are substantial barriers technical, analytical and communication barriers.	Sentiment analysis	2
3	U.S.	2010	Content and motivations for sharing health-related activity on social media	Positive, negative and neutral sentiment towards the activity are reportable within message content	Activity actualization and sentiment	4
4	U.K.	2014	Hashtags contribution to relevance	Twitter facilitates one-to-many, asynchronous communication	Exploratory Analysis	21
5	U.S.	2013	The Use of Social Media at Commercial Service Airports	Findings are consistent with public management techniques; cost-efficiency, reduces expenditures associated with traditional marketing, engages audiences.	Exploratory Analysis	9

literature review focuses on two major topics: 1) social media mining in transportation research, and 2) motivation behind bike commuting.

2.1. Social media mining in transportation research

Social media data mining is a growing field where analysts dissect social media posts using varying text mining techniques and is gaining popularity due to its cost effectiveness, accessibility and anonymity [2,3]. Some techniques involve a selection of keywords while others plan a campaign hashtag which is later used to isolate related Twitter posts. Recently, several studies have conducted sentiment analyses through text mining on topics ranging from; transportation planning [2], health related activity [4], and commuter bike sharing programs [3]. Researchers all over the globe wish to examine characteristics of tweets to understand electronic word of mouth [2,5]. Some study analytic tools and methods to see what findings may emerge from their use [3]. Others hypothesize there are existing psychological, social and behavior related theories that may be applied in a social media context [6,7].

Twitter network analyses take different forms. However, examining retweets and mentions–the back-and-forth conversational element of Twitter–has proven challenging [2,3]. Social media in the transportation sector frequently cites published research on best practices and the lessons learned from analyzing these public communication platforms.

Health researchers from Rutgers University analyzed Twitter posts to identify the type of content and motivations for sharing health-related activity on social media [4]. Teodoro's et al. method consisted of an Application Program Interface (API) search using exercise, diet, and weight related keywords–like the selection of bike commuting related hashtags to conduct a qualitative analysis of Twitter posts activity actualization and sentiment in this study [4].

The researchers are aware of the multitude of social media micro-blogging platforms from which one could learn from, including Facebook, Snapchat, Instagram, and more. However, as this paper documents a technique for analyzing Twitter posts, Twitter is the focus of discussion in this section. Twitter's mission is “to give everyone the power to create and share ideas and information instantly, without barriers” [8]. The transportation sector has adopted social media for many purposes, including, but not limited to: traffic safety campaigns, survey promotion, and traffic management [9]. Many national, state, and local agencies use social media to promote bike commuting, raise awareness, and promote safety. For example, the National Highway Traffic Safety Administration (NHTSA) supports other organizations' social media efforts by providing electronic toolkits that contain graphics, images, statistics, and text making it convenient for social media managers to repurpose the content for their own channels [10]. Majumdar examined to gauge the extent of social media use in transportation planning among local government agencies and identify major challenges [11]. Bao et al. investigated how to incorporate human activity information in spatial analysis of crashes in urban areas using Twitter check-in data [12]. In their study, Garcia-Palomares showed that social network data can be used to improve our understanding of the link between land use and urban dynamics [13]. The findings from Flores and Rezende study showed that Twitter enhanced transparency and strengthens bonds between local government and citizens [14]. Delbosc and Mokhtarian developed a purpose-designed survey to identify the association between physical and virtual social interaction [15]. In a way to explore the relationship with face-to-face social interaction, the developed multiple regression and structural equation modeling showed that more frequent ‘virtual’ interaction was associated with the more frequent face-to-face interaction. Nisar and Prabhakar examined the impact of Twitter content on users' train journeys and how train providers' message framing moderates these associations [16].

Table 2
Studies on bike commuting and motivations.

No.	Location	Study period	Research hypothesis	Findings	Method	Ref
1	U.S.	2015	Social media mining to identify attitude of Capital Bike Share of D.C.	Results revealed higher positive sentiments towards the current system.	Sentiment Analysis	3
2	U.K.	2015	A review of evidence on impacts and processes of implementation and operation of bike sharing	A positive cycling culture exists; evidence on users and usage of bike sharing.	Exploratory Analysis	18
3	U.S.	2012	Bike to Work Day analysis	Identified reasons for participating in the event, influence of event on participant, and demographics across behavior groups (year-round commuter, frequent biker, etc.).	Survey	19
4	U.S.	2015	Bike commuters community practice	Bike commuters was identified as a collective effort, but was occurred without organization.	Exploratory Analysis	6
5	U.K.	2009	Community, trust and social influence among commuter cyclists in the UK	The process of sharing information could perform not only a functional role in diffusing instrumental travel information, but also a social one.	Survey and observational study	7

Please cite this article as: S. Das, et al., Extracting patterns from Twitter to promote biking, IATSS Research (2018), https://doi.org/10.1016/j.iatssr.2018.09.002

The findings showed that train operators use tweets to understand public sentiment and expand information about schedules and notices. Kim et al. study examined two key topics: social networks and activity-travel behavior, and social influence and travel decisions [17]. This study classified models, summarized empirical findings and discussed important issues that require future research.

Table 1 lists several key studies on transportation engineering related social media content analysis.

2.2. Motivation behind bike commuting

Several studies have investigated motivations behind bike commuting [6–8], but few have utilized social media data mining [3]. Das et al. performed a sentiment analysis on the Capital Bike Share of Washington D.C. by mining tweets by hashtag, #bikeshare, and mention, @bikeshare [3]. Findings suggest a higher positive sentiment towards the Capital Bike Share system and positive cycling culture, which is an important factor to have in place to sustain bike sharing [4,8]. Ricci et al. document findings that support Das et al. data mining methods, showing it is worthwhile to analyze social media posts to understand the public's view on bike commuting [3,18].

A prime example of how Twitter functions to promote community relations and outreach in the bike community is the social media outreach surrounding National Bike to Work Month/Week/Day. Traditional participant surveys have taken place to understand the motivations of bike commuters on Bike-to-Work Day (BTWD). Piatkowski et al. identified barriers to increased commuter cycling and cyclist behaviors through online surveying pre- and post-BTWD and categorized results by self-reported typical bicycling behavior (i.e. year-round commuter, only on BTWD, etc.), finding that BTWD motivates and impacts different groups in different ways. Results show over half of the participants who only bike on BTWD participate for fun, while over half of the year-round commuters participate to raise awareness [19]. A social media analysis of BTWD participants has not been conducted.

Individuals and companies who make up the bicycle commuter community take to social media for personal and/or professional reasons, uncertain whether their posts have influence on their followers. Bartle et al. studied commuter cyclists' information sharing habits to understand their social influence and found that information sharing serves as a social role that builds trust and reinforces a positive view of cycling. The same could be said about information sharing over social media. Cyclists' word-of-mouth offered encouragement to those new to cycling and suggests ‘user-generated’ information may potentially be used as a tool for promoting sustainable travel [7]. Table 2 lists studies that explored the motivation behind bike commuting.

3. Methodology

Recent social media engagement statistics show 1.71 billion monthly active users of Facebook and 313 million Twitter accounts worldwide [20]. However, these users are not evenly spread across the world. A review of relevant literature was conducted to provide the foundation for this study. In addition, data mining of the well-known social media network, Twitter, was conducted to understand the bike commuter network. To understand the bike commuter community through Twitter post analyses over time, researchers use a technique for text mining by hashtag. A hashtag is denoted by a word with a preceding “#” symbol (e.g., #biketowork). Hashtags are generally used before a relevant keyword or phrase, with no spaces, in tweets to categorize those tweets and help users track content and updates relevant to the posted topic [21].

The pragmatics of hashtags have evolved since 2009 when Twitter formally adopted hashtags as a feature of the site [22]. Twitter users must now be aware of contextual assumptions that may need to be made both explicitly and implicitly [21]. This revelation adds some uncertainty to a formal analysis of Tweets, however, some may argue that with a 140 character limit to begin with, a clear interpretation of Twitter posts is already a challenge.

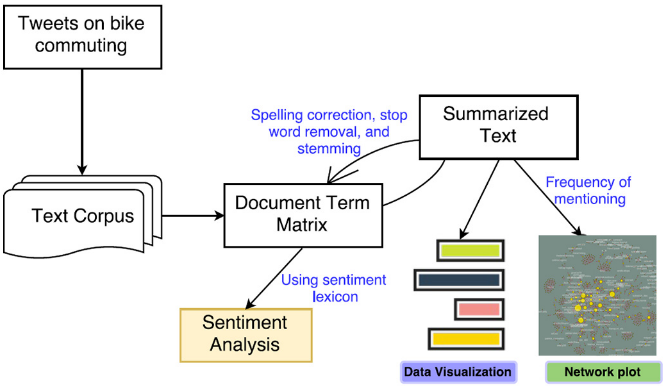

Fig. 1. Key procedures in step 2 and step 3.

Individual persons, business and organizations utilize social media for a variety of purposes. Incorporating trending hashtags into a social media presence is a way to join the conversation and create new dialogue [22]. There are drawbacks to using hashtags; one being spammers who use a trending hashtag in a post that has nothing or little to do with the topic. This study applied different natural language processing (NLP) techniques to accomplish the research goals and attempt to identify non-spam tweets.

3.1. Twitter mining framework

Bicyclists can use Twitter to express their experiences and challenges without any bias. The high variance of the information that propagates through Twitter is real-time and it makes it a key player in understanding people's interactions and opinions with minimal efforts. However, collecting and processing data from Twitter require some technical expertise. This study developed a framework for using Twitter as a tool for understanding the public interactions regarding biking. The key steps of this framework are described below:

1. Step 1. Data Collection. To identify bike commuting related tweets on Twitter, researchers used a selection of bike commuting related hashtags. The search terms used for the related hashtags are: #biketowork, #bikecommute, #bikecommuting, #bicycletowork, #bicyclecommute, #bicyclecommuting, #bike2work, #cycletowork, #cyclecommute, #cyclecommuting, #biketocampus, #biketooffice, #biketoschool, #bicycletocampus, #bicycletooffice, #bicycletoschool, #cycletocampus, #cycletooffice, #cycletoschool, #bike2campus, #bike2office, #bike2school, #bicycle2campus, #bicycle2office, #bicycle2school, #cycle2campus, #cycle2office, and #cycletoschool. The research team used two open source R packages “twitteR” and “tm” to collect data from Twitter and perform the text

Table 3
Chart on the percentage of tweet frequencies per month during different years.

January	3.2	4.6	4.9	5.4	6.1	4.6	6.0	7.1
February	2.6	4.8	5.5	5.1	6.0	4.2	5.1	6.5
March	3.2	7.1	8.1	8.5	6.9	5.3	6.3	7.3
April	6.3	7.2	7.6	9.1	9.0	7.5	8.4	8.5
May	25.3	24.1	22.4	28.0	21.9	23.7	19.5	19.7
June	9.3	10.1	12.9	8.6	12.9	11.5	9.4	9.8
July	7.4	6.9	6.8	5.6	6.4	7.5	5.6	5.6
August	6.8	9.4	9.5	5.9	7.0	7.1	5.2	6.5
September	9.3	8.8	6.2	6.9	7.3	10.3	11.2	8.4
October	11.1	8.3	6.0	7.0	6.7	8.2	11.2	6.5
November	8.5	8.1	5.4	5.2	5.7	5.7	6.7	8.2
December	6.9	6,2	4.8	4.7	4.2	4.5	5.4	5.8
	2009	2010	2011	2012	2013	2014	2015	2016

Table 4
Chart on the percentage of tweet frequencies per month during different days of the week.

January	6.3	6.8	5.8	5.7	5.9	5.6	4.7
February	6.2	6.8	5.5	5.6	5.3	4.8	4.3
March	8.4	8.8	7.0	7.6	6.0	6.6	5.6
April	9.7	8.8	8.6	9.4	8.5	7.3	7.3
May	14.0	14.0	20.6	17.9	18.5	24.1	33.4
June	8.8	7.4	10.3	11.0	13.7	9.6	9.0
July	6.2	5.2	6.3	6.4	7.1	6.3	5.6
August	7.4	7.6	7.2	7.3	6.7	6.1	6.1
September	9.0	9.0	7.2	8.6	9.3	11.4	7.2
October	8.9	9.3	8.1	8.2	7.8	8.0	7.0
November	8.5	9.3	8.0	7.2	6.2	5.4	5.5
December	6.7	7.2	5.5	5.1	5.1	4.9	4.2
	Satur	Sun	Mon	Tues	Wednes	Thurs	Fri

mining [23,24]. Twitter currently implements open standard for authorization (OAuth) to provide authorized access to the developers [25]. The developers can collect data by obtaining an access token. The historical data prior to 2015 were collected by using advanced search option in Twitter [26]. Related tweets (by using the search terms) were collected from January 2009 to December 2016. In 2009, only 1048 tweets were identified as using a hashtag related to bike commuting. This number in 2016 is 16,897 (an approximate 1500% increase). The initial size of the collected tweets was around 87,000. Researchers determined that some Twitter accounts included in the original data are bot accounts and generate tweets for company campaigns. These tweets were deleted due to redundancy and irrelevance. For the final analysis, 80,563 relevant tweets were analyzed.

2. Step 2. Data Cleaning. The collected tweets usually contain redundant data (for example, emoticons or weblinks). This study follows some simple steps to perform the data cleaning: 1) delete redundant and spam tweets to get relevant tweets; 2) consider each unique tweet as a document; 3) delete redundant and stop words (e.g., punctuation, numbers, auxiliary verbs, etc.); 4) determine sparsity (sparsity indicates the threshold of relative document frequency) to remove common words present in all documents. This study uses the sparsity threshold as 0.95;
3. Step 3. Text Mining. The most common use of text mining applications are term frequency plot and word cloud. This study conducted three text mining tasks: 1) frequency of tweet analysis by year and days of week; 2) perform overall term frequency analysis; and 3) network analysis to identify the association between Twitter users and followers (see Fig. 1 for Step 2 and Step 3).
4. Step 4. Context Mining. Providing context to the text is another important task. This study performed sentiment analysis and polarity scores to extract the knowledge from the hashtag related interactions of Twitter users.

4. Results and discussions

Researchers conduct a variety of analyses and data visualizations, including heat charts and a chord diagram to identify beneficial trends and understand the interactions between Twitter users in the bike commuting community.

4.1. Text mining

Upon gathering bike commuter data via hashtag new insights may be gathered using heat charts. Of the 80,000 plus tweets from the 2009 to 2016 year span, researchers chose to categorize hashtag activity by month. Table 3 represents Twitter microblogging conversations about bike commuting in the form of a heat chart. The heat chart shows the overall percentage of bike commuter related tweets occurring each year are higher in May. The League of American Bicyclists established May as 'National Bike Month' and biking is celebrated in communities from coast to coast. Established in 1956, National Bike Month helps to showcase the benefits of biking so that more people would be interested in biking. The League of American Bicyclists and other supporters of bike month strongly encourage bicyclists to post on social media in May. Social media toolkits with facts, figures, and graphics have been developed to help people "spread the word about the joys and benefits of bicycling" [27].

Twitter data lends itself to other time/date categories by which one may gather online community insights. For example, Table 4 shows bike commuter related Twitter activity by month and day of the week. The frequency count shows that bike commuting related tweets are posted mostly on weekdays (around 90% of the total tweets). Of all the days of the week, Friday is the most popular day in May for bike commuter community posts to Twitter.

Multiple inferences may be made as to why more or fewer bike commuters post on social media in certain seasons, months, or days of the week. For the purpose of this study, it is important to note that heat charts, such as Tables 3 and 4 could be useful in future analyses.

Text mining gives insight into community conversations and topic trends. Text mining is a method used when qualitative data needs to be analyzed; for example, Twitter posts. Text mining results in the most used words of the data set. For purposes of this study, the top thirty words were identified, and their frequency is shown across the 2009 to 2016 time span in Fig. 2.

The most frequent term is the most common measure in understanding the hidden trend of unstructured text data. The word "day" was used most frequently over the years, as shown by the dark orange displayed on the heatmap. Other popular words used over the years are "morning," "today," and "week." Bike commuters tend to tweet about the time in which they commute or when. Additionally, positive terms like "beautiful," "great," "nice," "love," "happy," and "good" are found in the most frequent words. The frequencies were higher in the recent years. Weather places a significant role in bike commuting.

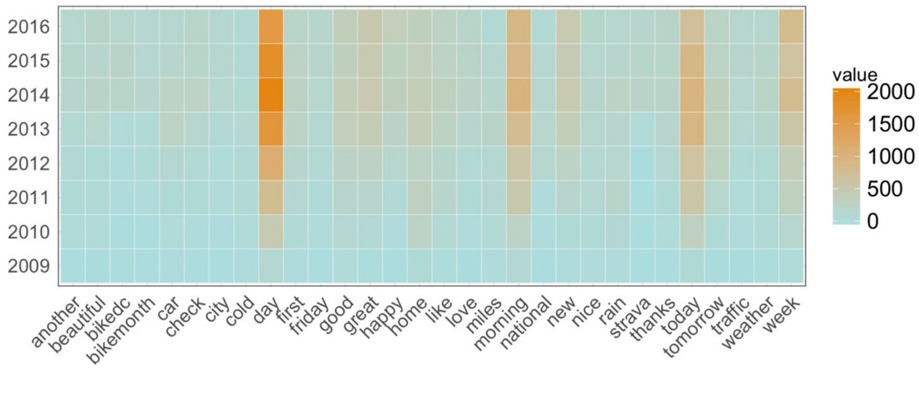

Fig. 2. Top 30 most frequent words.

Please cite this article as: S. Das, et al., Extracting patterns from Twitter to promote biking, IATSS Research (2018), https://doi.org/10.1016/j.iatssr.2018.09.002

Weather related words are also visible in the list of top 30 words. The words are “cold,” “weather,” and “rain”. The presence of the word ‘bikemonth’ in the list indicates that people tweet more in this month and it contributes significantly to the complete set of text data.

In general, a network is a collection of people or entities that connect or interact due to some standard criteria or situation. In a Twitter network, the Twitter handles are the people or entity, and the connections are followers. The current network is based on the standard criteria that involve usage of bike-commuting hashtags and mention someone in the same tweet. The purpose is to identify the influential handles in the Twitter network and their influence patterns. Network analysis through a chord diagram helps to understand the dynamics of the network, for example; information sharing between Twitter handles, expansion, and the cross relationship between different handles. In addition, network analysis can help obtain a holistic view of information distribution and it allows quantifications of an account's influence. The final dataset reveals 26,102 unique interactions. Distributions of these communications are below:

- Single interaction (20,471; 78% of total interactions);
- Two interactions (1424; 5% of total interactions);
- Three interactions (309; 1% of total interactions);
- Four interactions (120; <1% of total interactions); and
- Five or more interactions (176; <1% of total interactions).

A chord diagram (shown in Fig. 3) visualizes the interrelationships between Twitter handles/users that used bike commuting hashtags.

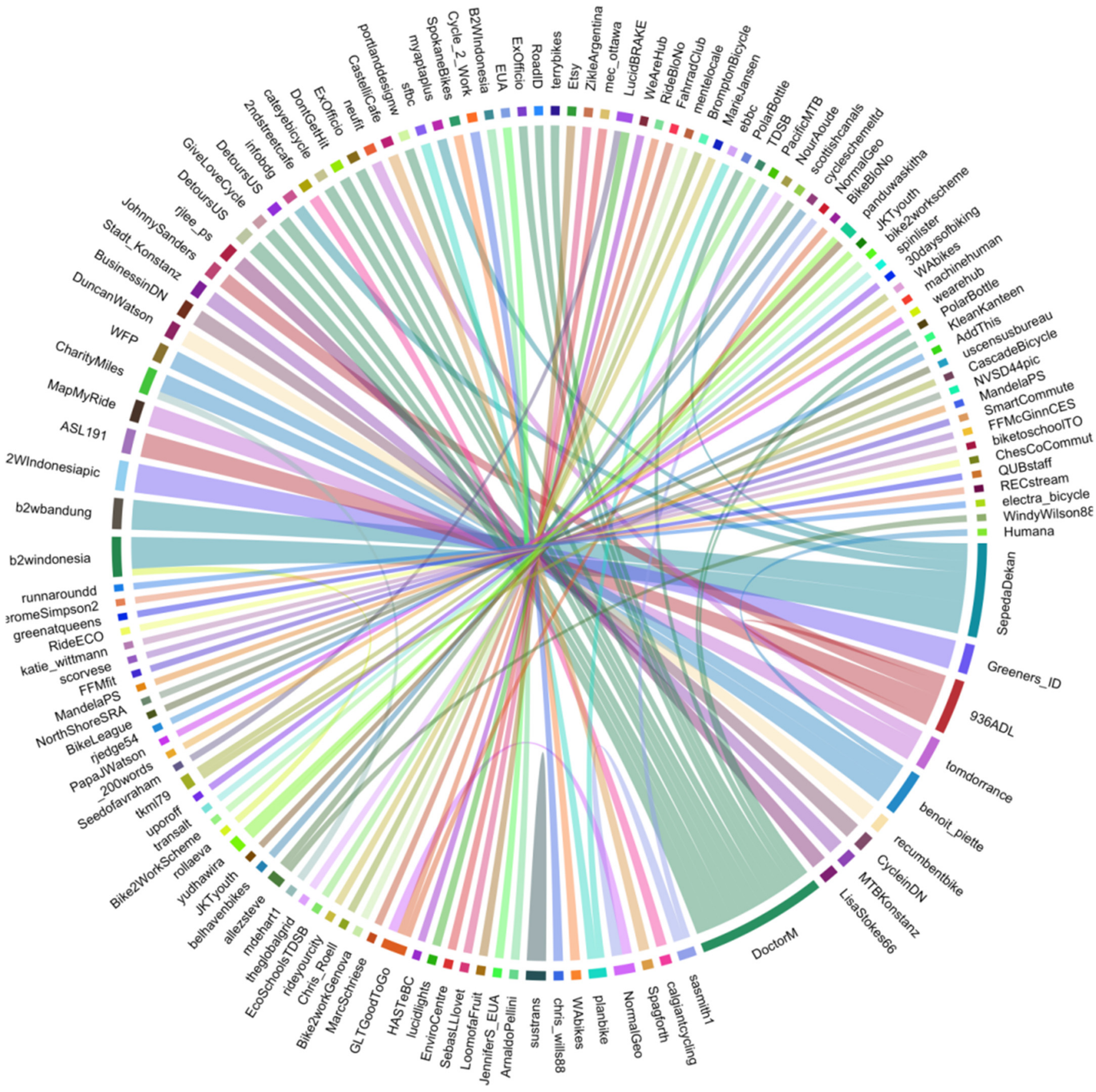

Fig. 3. Chord diagram of Twitter mention frequencies ($n > 4$).

The acquaintances between entities are used in displaying commonality of information or interest. Fig. 3 is helpful to compare the similarities and patterns within a dataset. In this figure, nodes are arranged in a circle, with the relationships between points connected to each other with either arcs or curves. Values, assigned to each connection, are represented proportionally by the size of each arc. The color is used in grouping the data into different categories that aid in making comparisons and distinguishing groups. To reduce the clutter, the authors only plotted interactions with frequencies above four.

This graphic indicates that some of the key handles (SepedaDekan, DoctorM, and 936ADL) are significant in networking. Twitter handles that lie on many short paths have considerable control over information diffusion in the network. This property is captured in a metric called betweenness centrality. Vertices with a high betweenness centrality lie on many of the shortest paths between the other vertices in the network. However, the overall network distribution does not identify centrality towards any common Twitter handle.

4.2. Context mining

4.2.1. Sentiment analysis

Studies on sentiment analysis or opinion mining have become exceptionally popular in recent years due to easy data access and real-world application. This analysis examines people's perception of a product, practice, or information by using an autonomous text mining algorithm instead of a manual method to simplify the process and reduce the level of difficulty for researchers. Sentiment analysis usually uses sentiment lexicon to provide sentiment scores on the generated corpus (a textual body clustered by required class or cluster).

The analysis focuses on individual sentence targets to determine whether a sentence expresses an opinion or not (often called subjectivity classification), and if so, whether the opinion is positive or negative (called sentence-level sentiment classification) [28].

Let an opinionated document be t, which can be a tweet that evaluates or expresses on a subject or a group of subjects. In the most general case, t consists of a sequence of words or sentences $t = \langle w_1, w_2, \ldots., w_n \rangle$. Definition of a sentiment passage on a feature is as follows – "A sentiment on a feature f of an object o evaluated in t is a group of consecutive words or sentences in t that expresses a positive or negative opinion on f" [28]. Additionally, sentiments also contain subjectivity. A subjective sentence expresses some personal feelings or beliefs. Document level sentiment classification involves a definite task with assumptions. These are stated below:

- *Task:* Given a set of opinionated tweets t, it determines whether each tweet $t \in T$ expresses a positive/negative/uncertain/litigious sentiment on an object. Given an opinionated document t that comments on an object o, determine the orientation oo of the opinion expressed on o, that is, discover the opinion orientation oo on feature f in the quintuple (o, f, so, h, p), where $f = o$ and $h, p,$ *and* o are assumed to be known or irrelevant.
- *Assumption:* The opinionated tweet $t =$ expresses opinions on a single object o and the opinions are from a single opinion holder h.

For transportation engineering, it is important to develop a domain specific lexicon, which is out of the scope of the current study. This study used three popular sentiment lexicons to perform this analysis. The lexicons are:

- Sentiment lexicon developed by Saif Mohammad and Peter Turney [29]
- Sentiment lexicon developed by Bing Liu [30]
- Sentiment lexicon developed by Loughran and McDonald [31]

Researched examined and partially modified the lexicons mentioned above were to develop a more relevant bike commuting related sentiment lexicon. For example, for crash analysis, the term 'risk' is a negative sentiment. For bike commuting, it occasionally related to risk to bike or not bike due to either time or weather constraint. This study considers 'risk' as an uncertainty related term rather than 'negative' term. Due to the conversational style of Twitter and 140 character limitation, users implicitly post to Twitter [20]. This makes interpretation of tweet sentiment difficult.

Table 5 lists sample bike commuting tweets to provide some understanding of the lexicon in use and assess the tweet author's perceptions.

Instead of developing an aggregate level sentiment score by considering all tweets as a corpus, this study used a disaggregate approach in discovering sentiments of the bike commuters. By a using sentiment score algorithm, words/terms were tagged in four sentiment classifications: 1) positive, 2) negative, 3) uncertain, and 4) litigious.

The relative frequency of positive tweeting is around 1.5% while the percentage of negative tweeting is around 0.5% (as shown in Fig. 4; legend colors are slightly changed due to overlapping). Positive and negative tweets have been on the upward trend since 2014.

To extract more knowledge of bike commuters' perceptions, researchers used text mining techniques to determine the most frequently used words for each sentiment classification (shown in Fig. 5). The quantities of positive tweets are higher than negative tweets. The median of the top 15 positive words is 490 [Inter quantile range (IQR): 250–1030]. For the top 15 negative words, the median is 120 [IQR: 90–190]. Uncertain words have a median of 90 [IQR: 70–220]. Additionally, litigious words (terms associated with legal or law related terminologies; for example: strict helmet law enforcement in a city or town can be intervened in the tweets) have the lowest median [Median: 18, IQR:10–30]. These statistics indicate that people express positive notions towards bike commuting when they post on Twitter. These findings are similar to earlier bike commuter studies in that bike commuter behavior and culture generally leans positively [3,6,7]. Social media generates information that could be analyzed to influence decision makers, bike commuting activists, transportation planners, and others alike [2,3]. Positive, negative, uncertain and litigious aspects to bike commuting mentioned on Twitter are presented in Fig. 5. The top most frequent words paired with negative sentiment are: "challenge," "bad," "bridge," "late," and "break." The top most frequent words paired with positive sentiment are: "great," "good," "happy," "beautiful," and "better."

Table 5
Sample bike commuting tweets and perception.

No.	Sample tweets	Perception
1	Great time yesterday at #Bike2Work #EarthFest. Wonderful to see our #MesaAz VIP's joining the cause! More #EarthDay	Positive experience sharing
2	It's great exercise too! #AdeptPR #BikeRide #Bike2Work	Motivational
3	Happy #CycleToWorkDay everybody! I love a good #Bike2work. Great to see lots of people walking/cycling every morning in recent weeks.	Information
4	Happy first day of #Bike2Work Week! <pic>	Picture sharing and first day of bike commuting
5	Headwinds are the worst #BikeCommute	Constraints in bike commuting
6	One of my worst PDX #BikeCommute home ever. Winds almost knocking me over. Glad there was #CraftBeer at home!	Constraints in bike commuting
7	Finding bike parking in this city is the WORST!! More #bikeracks pls #nyc @NYCMayorsOffice @transalt #bikecommute	Constraints in bike commuting
8	Missed the #DRagonA launch because #bikecommute and dog sitter responsibilities	Slower commute
9	Little risk of getting hot and sweaty this morning #bikecommute	Not associated with safety risk
10	Calculated risks don't always pay off. Light drizzle when I left the house #bikecommute	Not associated with safety risk

Please cite this article as: S. Das, et al., Extracting patterns from Twitter to promote biking, IATSS Research (2018), https://doi.org/10.1016/j.iatssr.2018.09.002

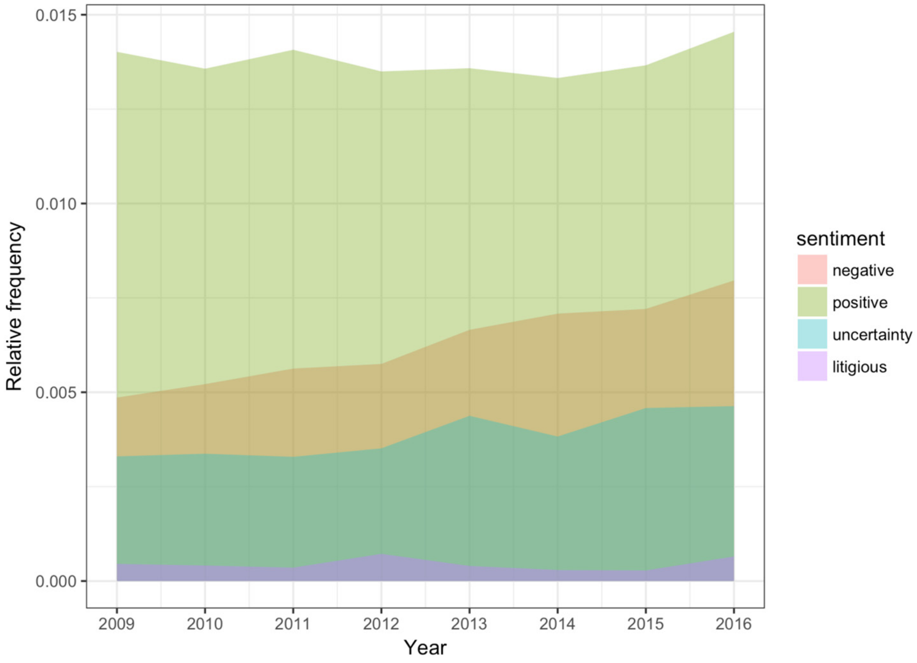

Fig. 4. Relative frequencies of different sentiments.

4.2.2. Polarity

Measuring polarity is a form of context mining. To develop a yearly polarity pattern, open source R package 'qdap' was used [32]. The polarity score determines the measure of polarity at the sentence level or a group of sentences based on certain clusters. The scores can be reweighted based on a revised lexicon or weightage of trigger words (words with the higher association for a certain context). For example, the word pair 'bad weather' may have different contexts for different scenarios. It may have the higher association with failure to make a bike trip and its associated sentiments. On the other hand, this might be associated with a crash occurrence for a motorist. The consequences and sentiments of a crash and failure to make a trip are different. In this study, each polarized word is weighted based on the sentiment lexicons

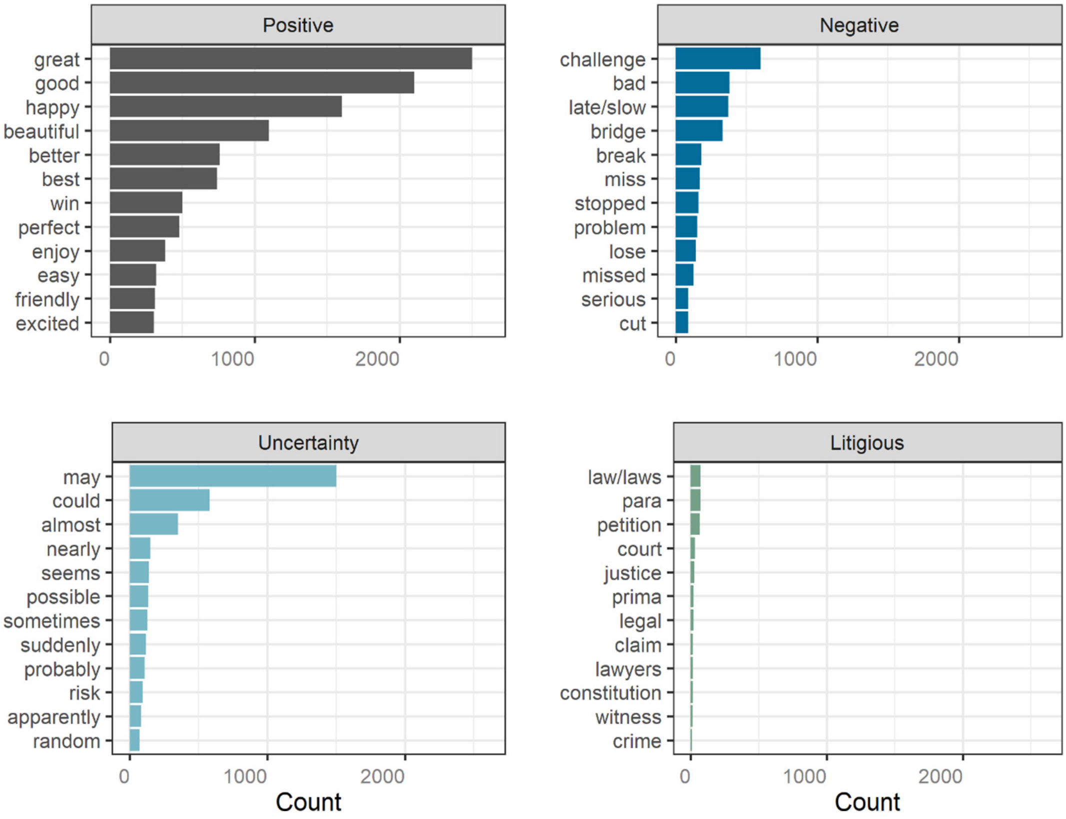

Fig. 5. Fifteen most frequent words for each sentiment classification.

Please cite this article as: S. Das, et al., Extracting patterns from Twitter to promote biking, IATSS Research (2018), https://doi.org/10.1016/j.iatssr.2018.09.002

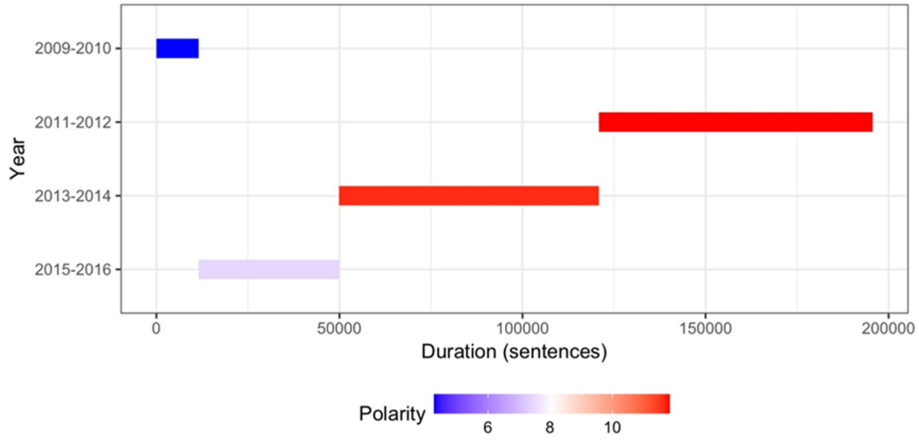

Fig. 6. Polarity patterns for different years.

used for the sentiment analysis. Fig. 6 shows the polarity patterns of the tweets divided into four different temporal groups. The x-axis in Fig. 6 indicates the duration of the polarities (how long similar polarity can sustain in the documents). The trend of the positive polarity changes as time advances. Polarities that are more positive were seen in the middle years (2011–2014). Less positive polarities are seen in both 2009–2010, and 2015–2016. This aggregate polarity plot requires more disaggregate level investigation to identify the change of the polarity patterns over the years, however, this task is out of the scope for the current study.

5. Conclusions

In the recent years, many cities have invested in bicycle-friendly programs and infrastructure. This trend increases the demand for and relevance of bicycle data as well as data related to people's perception of bicycle commuting. Local governments and planning agencies are interested in understanding the factors that triggered bicycle commuting. There is a need for understanding the interaction and sentiment of bike users for a broad perspective. Social media offers bike riders a free, interactive and real-time platform for expressing their challenges and experiences, which can be viewed by millions of other users. This study has two unique contributions: 1) it developed a framework of collecting and analyzing biking associated texts from Twitter by performing text mining and context mining, and 2) it performed a comprehensive sentiment analysis using approximately 10 million words from 80,00 tweets. Some specific findings of this study are the following:

- There is a surge of bike commuting tweets in the recent years (approximated 1500% increase of tweets in seven years during 2009–2016). This is in line with the findings on ACS study, which showed that the number of U.S. bicycle commuters increased from about 488,000 in 2000 to about 786,000 (around 60% increase) from 2008 to 2012.
- The text mining on the tweet data showed several specific findings. Weekday tweets are dominating in terms of frequency (around 90% of the total tweets). Due to the celebration of bike month in May, the higher number of tweets are seen in May. According to a recent report, “Bike to Work events happened in all 52 of the largest U.S. cities in 2012. In 2014, thousands of residents participated in cities like Denver (30,000), Chicago (12,000) and Boston (10,000) —with each community seeing sizable gains in just one year [27].” The most frequent words show trends towards temporal information, weather, traffic, bike month, and positive experiences of biking. The network analysis based on the Twitter mentioning does not show any central tendency. The information sharing patterns are scattered and sparse. It indicates the bike events and popular biking movements are mostly locality specific.
- The context mining generates the general contexts of the tweets associated with bike riding experience. The current study did not explore the local contexts due to the unavailability of the geotags of the tweets. However, the general contexts are most likely to be universal. The relative frequency of positive tweeting is around three times higher than negative tweets. In general, people express more positive tweets while using bike-commuting hashtags. The negative tweets indicate potential constraints in bike commuting. The problems include slower speed, headwind, wind, rain, water on road, crime, and bad weather. The terms associated with negative and uncertain tweets are in line with the finding of Ahmed et al. study, which showed that half of the variations in bicyclist volume can be explained by the changes in weather parameters [33]. The polarity is somewhat positive in context when temporal patterns are considered.

In the recent years, there is a surge of text mining related transportation engineering studies [34–40]. The current study develops a framework on the application of social media mining in understanding the public perception of biking and presents several opportunities for future research. Spatial segregation of the tweets would be beneficial in identifying localized contributing factors, which was not done in the current study. The network analysis was conducted on a very high level. Network patterns in small groups require further exploration and future studies. The aggregate level polarity levels show discontinuity in positive polarity trends. A disaggregate level analysis could shed more light on the polarity patterns. MacEachren et al. argue that the majority of text mining studies focused on data extraction and text categorization but conducted limited efforts in transforming the findings into actionable knowledge [41]. It is important to relate the findings to knowledge application. The findings from this study can be used as important resources for transportation planners, community activists, and policy makers to assess important factors and trends. This method develops a framework to discover influence and motivation parameters in bike commuting and is replicable for other transportation related topics. In conclusion, this study has shown that interpreting unstructured data is attainable through text mining and context mining.

Acknowledgements

The authors appreciate the assistance provided by Maria C. Rodriguez and Anandi Dutta. The authors also like to acknowledge the excellent comments and suggestions provided by three anonymous reviewers.

References

[1] B. McKenzie, Modes less traveled: commuting by bicycle and walking in the United States: 2008–2012, American Community Survey Reports, ACS-26, U.S. Census Bureau, Washington, DC, 2014.

[2] J. Evans-Cowley, G. Griffin, Microparticipation with social media for community engagement in transportation planning, J. Transp. Res. Board (2012) 90–98, https://doi.org/10.3141/2307-10 No. 2307.

[3] S. Das, X. Sun, A. Dutta, Investigating user ridership sentiments for bike sharing programs, J. Transport. Technol. 5 (2015) 69–75, https://doi.org/10.4236/jtts.2015.52007.

[4] R. Teodoro, M. Naaman, Fitter with Twitter: Understanding Personal Health and Fitness Activity in Social Media, *ICWSM*, 2013 611–620.

[5] E. Lahuerta-Otero, R. Cordero-Gutiérrez, Looking for the Perfect Tweet. The use of Data Mining Techniques to find Influencers on Twitter, Comput. Hum. Behav. 64 (2016) 575–583, https://doi.org/10.1016/j.chb.2016.07.035.

[6] E.D. Wilhoit, L.G. Kisselburgh, Collective Action without Organization: the Material Constitution of Bike Commuters as Collective, Organ. Stud. 36 (5) (2015) 573–592, https://doi.org/10.1177/0170840614556916.

[7] C. Bartle, E. Avineri, K. Chatterjee, Online Information-sharing: a Qualitative Analysis of Community, Trust and Social Influences Amongst Commuter Cyclists in the UK, Transportation Res. Part F 16 (2013) 60–72.

[8] About Twitter, https://about.twitter.com/company Date accessed June 5, 2017.

[9] C.L. Stambaugh, Social Media and primary Commercial Service Airports, Transp. Res. Rec. (2013) 76–86, https://doi.org/10.3141/2325-08 No. 2325.

[10] United States Department of Transportation National Highway Traffic Safety Administration, Traffic Safety Marketing, https://www.trafficsafetymarketing.gov/ Accessed on July 10, 2017.

[11] S. Majumdar, The case of public involvement in transportation planning using social media, Case Stud. Transp. Pol. 5 (1) (2017) 121–133.

[12] J. Bao, P. Liu, H. Yu, C. Xu, Incorporating twitter-based human activity information in spatial analysis of crashes in urban areas, Accid. Anal. Prev. 106 (2017) 358–369.

[13] J. García-Palomares, M. Salas-Olmedo, B. Moya-Gómez, A. Condeço-Melhorado, J. Gutiérrez, City dynamics through Twitter: Relationships between land use and spatiotemporal demographics, Cities 72 (2018) 310–319.

[14] C. Flores, D. Rezende, Twitter information for contributing to the strategic digital city: Towards citizens as co-managers, Telematics Inform. 35 (5) (2018) 1082–1096.

[15] A. Delbosc, P. Mokhtarian, Face to Facebook: the relationship between social media and social travel, Transp. Policy 68 (2018) 20–27.

[16] T. Nisar, G. Prabhakar, Trains and Twitter: firm generated content, consumer relationship management and message framing, Transp. Res. A Policy Pract. 113 (2018) 318–334.

[17] J. Kim, S. Rasouli, H. Timmermans, Social networks, social influence and activity-travel behaviour: a review of models and empirical evidence, Transp. Rev. (2017) 499–523.

[18] M. Ricci, Bike sharing: a review of evidence on impacts and processes of implementation and operation, Res.Transport. Business & Manag. 15 (2015) 28–38, https://doi.org/10.1016/j.rtbm.2015.03.003.

[19] D. Piatkowski, R. Bronson, W. Marshall, J. Krizek, Measuring the Impacts of Bike-to-Work Day events and Identifying Barriers to increased Commuter Cycling, J. Urban Plann. Developt. 141 (2015).

[20] Statista, *Social* Media Statistics *and Facts*, https://www.statista.com/topics/1164/social-networks/ Accessed on July 10, 2017.

[21] K. Scott, The Pragmatics of Hashtags: Inference and Conversational style on Twitter, J. Pragmat. 81 (2015) 8–20, https://doi.org/10.1016/j.pragma.2015.03.015.

[22] A. MacArthur, The History of Hashtags: Shedding some Light on the History of Hashtags and how We've Come to Use Them, http://twitter.about.com/od/Twitter-Hashtags/a/The-History-Of-Hashtags.htm Accessed on July 10, 2017.

[23] J. Gentry, twitteR: R Based Twitter Client. R package version 1.1.9, https://CRAN.R-project.org/package=twitteR 2015.

[24] I. Feinerer, K. Hornik, Tm: Text Mining Package. R package version 0.7-4, https://CRAN.R-project.org/package=tm 2018.

[25] Twitter Oauth, https://developer.twitter.com/en/docs/basics/authentication/overview/using-oauth Accessed July 27, 2018.

[26] Historical tweet Access, https://github.com/Jefferson-Henrique/GetOldTweets-java Accessed July 27, 2018.

[27] The League of American Bicyclists, Promotional Materials for National Bike Month, http://bikeleague.org/content/promotional-materials-0 Accessed on July 31, 2017.

[28] N. Indurkhya, F. Damerau, Handbook of Natural Language Processing, Chapman & Hall/CRC, 2010.

[29] S. Mohammad, P. Turney, Crowdsourcing a Word-Emotion Association Lexicon, Comput. Intell. 29 (3) (2013) 436–465.

[30] B. Liu, Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, 2012.

[31] T. Loughran, B. McDonald, Master Dictionary of Loughran and McDonald, https://www3.nd.edu/~mcdonald/Word_Lists_files/Documentation/Documentation_LoughranMcDonald_MasterDictionary.pdf Accessed on July 31, 2017.

[32] T.W. Rinker, Qdap: Quantitative Discourse Analysis Package, University at Buffalo, Buffalo, New York, 2013.

[33] F. Ahmed, G. Rose, C. Jacob, Impact of weather on commuter cyclist behavior and implications for climate change adaptation, ATRF 2010: 33rd Australasian Transport Research Forum, Australia, 2010.

[34] S. Das, L. Minjares-Kyle, K. Dixon, A. Palanisamy, A. Dutta, TRBAM: exploring knowledge management, research trends, and networks by social media mining, The Proceedings of Transportation Research Board 97th Annual Meeting, Washington D.C., January, 2018.

[35] F. Oliveira-Neto, L. Han, M. Jeong, Tracking large Trucks in Real Time with License Plate Recognition and Text-Mining Techniques, Transport. Res. Rec. 2121 (2009) 121–127.

[36] S. Das, A. Mudgal, A. Dutta, S. Geedipally, Vehicle Consumer Complaint Reports Involving Severe Incidents: Mining large Contingency Tables, Transportation Res. Rec. (2018) 1–11, https://doi.org/10.1177/0361198118788464.

[37] S. Das, B. Brimley, T. Lindheimer, A. Pan, Safety Impacts of Reduced Visibility in Inclement Weather, Final Report- ATLAS-2017-19, 2017.

[38] S. Das, K. Dixon, X. Sun, A. Dutta, M. Zupancich, Trends in Transportation Research: Exploring Content Analysis in Topics, Transportation Res. Rec. 2614 (2017) 27–38.

[39] R. Boyer, W. Scherer, M. Smith, Trends over two decades of Transportation Research: a Machine Learning Approach, Transportation Res. Rec. 2614 (2017) 1–9.

[40] S. Das, X. Sun, A. Dutta, Text mining and topic modeling on compendium papers from transportation research board annual meetings, Transportation Res. Rec. 2552 (2016) 48–56.

[41] A. MacEachren, A. Jaiswal, A. Robinson, S. Pezanowski, A.S.P. Mitra, X. Zhang, J. Blanford, Senseplace2: Geotwitter Analytics support for Situation Awareness, Presented at the IEEE Conference on Visual Analytics Science and Technology, Providence, RI, 2011.

Please cite this article as: S. Das, et al., Extracting patterns from Twitter to promote biking, IATSS Research (2018), https://doi.org/10.1016/j.iatssr.2018.09.002