Social Media Analysis as a Method of Exploring Cyclists' Attitudes

CAROLINE JAFFE*[§], JAN ANNE ANNEMA*[**], and BERT VAN WEE*[††]

**Transport and Logistics, Delft University of Technology, Jaffalaan 5, 2628 BX Delft, the Netherlands*

ABSTRACT This paper evaluates social media analysis, specifically with Twitter data, as a method of studying cyclists' attitude. We use a literature overview to design two case studies where machine-learning techniques are applied to Twitter data and use the studies' results to explore the benefits and challenges of using social media data. We conclude from our case studies that Twitter data—and social media data in general—offer some interesting advantages over traditional cycling data. The nature of the data allows researchers to address a new class of questions related to users' motivation, attitude and emotion. One of the biggest challenges of working with Twitter is understanding the representativeness and accuracy of the data. Unlike data obtained from more traditional methods, Twitter data does not include socio-demographic metadata and it can be difficult to ensure accuracy. We use the challenges encountered in our case studies to launch a discussion of the solutions, benefits, and possibilities of using Twitter data in cycling research.

Keywords: bicycling, bicycling attitudes, social media, Twitter, sentiment analysis, topic modeling

Acknowledgements: This work was supported by the Netherlands-America Foundation through a yearlong NAF-Fulbright research grant. The authors would also like to thank the Transport and Logistics group at Delft University of Technology and the Internet of Things group at Bell Labs Antwerp for their valuable input during two colloquia on this topic.

[§] Corresponding author: caroline.jaffe325@gmail.com, +31 (0) 68 17 39125
[**] J.A.Annema@tudelft.nl, +31 (0) 15 27 88912
[††] G.P.vanWee@TUDelft.nl, +31 (0) 15 27 81144

1. Introduction

Cycling receives growing recognition as a healthy, convenient, sustainable form of urban transportation with benefits—emission reductions, increased physical activity, reduced congestion and fuel use, flexible mobility—that may offset environmental and public health concerns (Shaheen et al., 2010). Despite increased efforts to make cycling an accessible and attractive option through government funding and infrastructure investments, cycling faces low rates worldwide and is not yet considered a viable form of daily transportation in most parts of the world (Bicycle and Pedestrian Provisions, 2014; Flegenheimer, 2013). Pucher and Buehler (2012) offer some statistics: the Netherlands has the highest cycling rate in the world, with 27% of all trips made by bicycle. In Denmark, the rate is 18%; in Germany, 10%. These countries have invested in longstanding research efforts and poured money into infrastructure and outreach for many years. On the other hand, the cycling rate in the U.S. stands around 1%. Even the best cycling cities in the U.S. have very low rates: Davis, Boulder and Portland have rates of 15.5%, 9.6% and 6.0%, respectively. All other U.S. cities have rates below 5%.

For cities to improve these rates and take advantage of the benefits of cycling, we must understand the factors that motivate bicyclists' choice to cycle or not. While some factors may seem intuitive, such as the preference for dry weather over rain, there are other factors that merit a more thorough and quantitative investigation. In particular, attitude towards cycling is an important—though poorly understood—factor that seems to affect cycling behavior. Heinen et al. (2010, pp. 83) write that "attitudes play a more significant role in mode choice than has so far been assumed," and suggests that "comprehensive research into commuting by bicycle should...focus on attitudes and people's social environments". A well-developed understanding of the different factors that influence cycling can inform policy decisions, infrastructure proposals and outreach programs.

Research in this field generally uses methods like questionnaires, counters, and to a lesser extent, interviews (Section 2.2). The aim of this paper is to evaluate social media analysis as a research method that can be used instead of or alongside more traditional methods of studying bicycling. While there has been growing interest in utilizing social media data in a number of different fields (Section 2.3), to the best of our knowledge it has not been used for travel behavior research. Our scientific contribution is that we evaluate and discuss the utility of a novel method: applying machine learning techniques to Twitter data related to cycling.

First, we conduct an overview of current cycling research and methodologies and projects that utilize social media analysis. Drawing on the advantages and disadvantages culled from this literature overview, we design and carry out two case studies using Twitter data. We are specifically interested in exploring the types of questions that can be asked of Twitter data, which we choose to use because it is most well suited to our research. Largely public, Twitter is a social network focused on sharing opinions, ideas, experiences and preferences. Relative to other types of social media, we found Twitter's unfiltered, lexically complex nature most appropriate for our research inquiries. In our case studies (Section 3), we apply machine-learning

techniques to Twitter data to build an understanding of the motivations, attitudes, and emotions surrounding bicycling in New York City (NYC) and seven other U.S. cities where cycling is a visible subject of public interest and debate. The result of our study is an in-depth discussion of the benefits, challenges and possibilities of social media analysis, focusing specifically on Twitter data.

The structure of the paper is as follows: in Section 2, we present a literature overview and discussion of current work with Twitter data; in Section 3, we carry out the two case studies based on lessons learned in Section 2 and describe our results; in Section 4, we offer an in-depth discussion of the potential of social media analysis; finally, we present our conclusions in Section 5.

2. Literature Overview

2.1 Factors That Impact Cycling

Developing a better understanding of the factors that influence cycling behavior, as well as the policies that can affect these factors, is an important first step towards increasing cycling mode share. In general, researchers agree that no single factor is entirely responsible for improving bicycle mode share; instead, it is likely a complex blend of factors. Heinen et al. (2010) carry out a thorough review of the factors that correlate with bicycling behavior, while Pucher and Buehler (2008) discuss how government policies impact some of these factors. Heinen et al.'s work is a comprehensive review that draws upon many other important findings, so we use their study to introduce a list of factors that influence cycling, presented in Table 1. They divide these factors into five categories: built environment, natural environment, socio-economic, practical and psychological. Psychological factors such as attitudes, social norms, environmental beliefs, and habits are of particular interest to us and are the focus of our paper. In general, they are poorly understood and Heinen et al. (pp. 74) recommend that "more specific research should be undertaken into bicycle commuting, using psychological theories".

Table 1: Individual-level factors that impact cycling

Category	Factor	Relationship with Cycling
Built environment	Urban form: network layout and neighborhood development	Dense layouts and heavily mixed-use development encourage cycling
	Infrastructure	Protected bike paths or lanes, lack of or reduced on-street car parking, continuous bike infrastructure, and good surface quality all encourage cycling
	Commuter amenities	Changing facilities, showers, and bike parking at work encourage cycling
Natural environment	Slope or hilliness	Flat environments preferred
	Environment attractiveness	Attractive environments preferred

	Climate or season (long-term)	Summer is the most common biking season
	Day-to-day weather (short-term)	Low chance of rain preferred
Socio-economic	Gender	Men cycle more often
	Age	Young people cycle more often
	Income	Unclear; cycling is cheap, which may attract lower-income populations, but those with high income tend to focus on their physical health more
	Car ownership	Those with cars cycle less
	Bike ownership	Those with bikes cycle more
	Household structure	Those with young families are less likely to cycle; individuals without children and students are more likely to cycle
	Level of physical activity	Those who are already physically active tend to cycle more
Practical	Cost	Low costs preferred
	Travel time	Shorter travel times preferred
	Safety	Safer conditions preferred
	Effort	Low-effort conditions preferred
Psychological	Attitudes	Those with a more positive attitude towards cycling are more likely to cycle themselves
	Social norms	Communities where cycling is more widely embraced are more likely to convert people to cycling or interest people in cycling
	Environmental beliefs	Those who care more about the environmental are more likely to cycle
	Habits	Not completely understood; people will stray from perfectly rational behavior to maintain a habit, whether the habit is cycling or not cycling

Source: Heinen et al. (2010)

While Heinen et al. (2010) focus on individual-level factors that correlate with cycling behavior, Pucher and Buehler (2008) offer a discussion of policies that can increase cycling mode share; while not all factors can be changed (e.g., the natural environment), others, such as the built environment, offer room for improvement. In their paper, they focus almost exclusively on governmental programs that have resulted in increased mode share in Germany, the Netherlands and Denmark; these countries experienced a sharp decrease in cycling during the 1950's to 1970's, but were able to revive mode share in the ensuing decades through national and municipal policies. Importantly, these countries are all industrialized, wealthy countries where many citizens own cars, but choose to cycle instead. Pucher and Buehler contrast these countries with the U.S. and the U.K., where cycling rates are extremely low. They attribute this difference to dramatic policy shifts in Germany, the Netherlands, and Denmark that prioritized

cycling and bicyclist safety, and restricted car use. These policies included planning, funding and coordinating improved bicycle infrastructure and facilities, developing cycle training and safety programs in schools, imposing traffic calming on residential streets, creating car-free zones in city centers, elevating the legal status of cyclists, and heavily taxing car use and gasoline. Most importantly, they note, is these countries' ability to deliver a coordinated, multi-faceted approach to increasing cycling mode share.

2.2 Traditional Methods of Bicycle Research

Just as research on bicycling behavior is still a developing field, so too is the field's methodology. Krizek et al. (2009) describe how the conceptual framework of cycling research depends on the researchers' background: public health researchers generally rely on psychology-based social ecological models, whereas transportation researchers generally use economic utility theory. In each case, collecting data on bicycling behavior poses a major challenge. The three main measurement methods used are: (1) surveying or interviewing people for the self-reported details of their travel behavior, (2) observing travel behavior manually or using counting technology, or (3) using sensing instrumentation on bodies or equipment to measure behavior (Troiano, 2005).

These methods appear in many important and widely cited pieces of bicycling research, such as Heinen's survey on commuter cycling behavior, Jacobsen's review of cycling injuries, which draws on survey data, and Dill's study on cycling infrastructure, which collects and analyzes GPS data (Heinen & Handy, 2012; Jacobsen, 2003; Dill, 2009). While this research offers valuable insights, these methods present distinct advantages and disadvantages, as summarized in Table 2. Self-reported measurements in surveys or interviews tend to be less accurate, because people have imprecise memories of their cycling behavior or unrealistic expectations of their future behavior and are generally bad at estimating trip distances (Agrawal & Schimek, 2007). Observing behavior can yield more accurate results, but offers no information about cyclists' motivation. Installing counters is accurate but expensive, as a single counter can cost $30,000 (Trimm, 2012). Instrumentation such as accelerometers or GPS systems is highly accurate but can be costly and presents its own set of data processing challenges. Collecting geographically diverse data can be prohibitively expensive and time-consuming. In certain communities, due to the small number of people who cycle regularly, it can be difficult or time-consuming to gather enough data to establish statistical significance (Krizek et al., 2009).

Most importantly, purely quantitative data collected through instrumentation offers no clues about cyclists' motivation or attitude, and thus cannot help researchers understand factors like attitude, habit and social expectation that may play a large role in determining users' mode choice.

Table 2: Advantages and disadvantages of traditional cycling research methods

Method Name	Advantages	Disadvantages
Surveying (online,	• Potential for information about why	• Accuracy

phone, or in-person)	people travel	• Difficulty of defining a trip • Limited amount of data
Observing behavior manually or using counting technology	• More accurate/ objective results than surveys	• No information about why people travel • Limited amount of data
Using sensing instrumentation on bodies or equipment to measure behavior	• Most accurate method for determining mileage	• Expensive • Limited amount of data • No information about cyclists' motivations

Source: Krizek et al. (2009); Agrawal & Schimek (2007); Trimm (2012)

2.3 Social Media Analysis with Twitter

Social media analysis, an area of growing research interest, can potentially overcome some of the disadvantages of these traditional methods, such as high costs and limited data. Social media is a general term for many types of Internet applications where people share ideas and information and interact with each other. All social media offers interesting research opportunities because it provides large-scale, dynamic data sets with information about relationships between users. In this paper, we focus specifically on Twitter, a social network and microblogging platform; our case studies utilize Twitter and our conclusions refer directly to Twitter. We choose to use Twitter because it is a network where users often publicly share their ideas, opinions, preferences and experiences. In this section, we offer an introduction to Twitter and examine several examples of current Twitter research to learn about some of the benefits and challenges of this type of research; these lessons are incorporated into the design of the case studies we present in Section 3.

2.3.1 Introduction to Twitter

Twitter is a microblogging service that was started in 2006. Users post "statuses" or "tweets" with a maximum of 140 characters that can consist of a status update, a link, a message or a photo or video. When users logon, Twitter prompts, "What's happening?" Twitter can be accessed via web browser or mobile, and other services—including Instagram, Foursquare, and Dropbox—can post to Twitter on a user's behalf. User profiles can represent individuals or other entities, such as a brand, institution, or musical group. Except for certain high-profile users, like celebrities or politicians, there is no verification system in place to ensure that users are who they say they are or are posting factually accurate information.

Two important Twitter concepts are the "hashtag", represented by the "#" character, and the "at", represented by the "@" symbol. Hashtags, such as #biking or #bike, associate that tweet with a particular concept or event; they "allow the content produced by many individuals to be aggregated into a public, topic-specific stream including all the tweets containing a given token" (Conover et al., 2013). The "@" symbol denotes a Twitter username or handle. Twitter handles not only represent one's unique identity on Twitter, but are also a way for users to communicate.

When composing tweets, users can mention another user's handle to address the tweet to that user and make sure that they see it.

As of August 2013, Twitter reported over 500 million tweets per day (Krikorian, 2013). Though the U.S. accounts for 28.9% of Twitter users overall, only about 11% of people in the U.S. use Twitter (www.alexa.com). While there is a degree of uncertainty about the socio-demographic distribution of Twitter users, we know that they tend to be young: 43% of Twitter users are 10-19 years old; 37% are 20-29 years old. Asian countries have the youngest users, pulling down the global average. The average age of U.S. Twitter users is 25 (Schoonderwoerd, 2013).

2.3.2 Current Work with Twitter

Twitter offers the research community a view of how news and ideas travel by "Electronic Word of Mouth" (Martínez-Cámara et al., 2014). Twitter data provides direct access to unfiltered opinions of users; additionally, it is easy, fast and cheap to collect large amounts of data. While all types of social media data have been used for analysis, Twitter has one of the most open and simple network structures.

In general, Twitter research consists of applied studies that use Twitter data to investigate a particular subject, and methodological studies that test or improve upon a particular algorithm used with Twitter data. Applied studies often use these methodologies to develop discipline-specific results. Examples of applied research include studies charting the spread of event-specific information (such as the Occupy Wall Street protests and the London riots), predicting election outcomes, and studying how happiness relates to an individual's location (Conover et al., 2013; Panagiotopoulos et al., 2012; Tumasjan et al., 2010; Frank et al., 2013). On the methodological side, there are papers examining Sentiment Analysis on Twitter data, studying different Twitter sampling methodologies, and discussing topic modeling with Twitter data (Pak & Paroubek, 2010; Gilbert, 2008; Hong & Davison, 2010). In each case, there are acknowledged benefits and drawbacks to the use of Twitter data analysis.

All studies find that Twitter can be used in a meaningful way to learn about the habits and beliefs of users in the study. In particular, Frank et al. (2013) find that Twitter can provide richer and more fine-grained location data than other methods of data collection. Tumasjan et al. (2010) find that data from Twitter can predict election results and political sentiment. Other studies find that Twitter's flexibility and dynamic nature make it useful for disseminating information, reaching new audiences and building relationships (Panagiotopoulos et al., 2012). All find that the immediate, unfiltered nature of Twitter data can reveal meaningful results that not only corroborate results from traditional studies, but also augment them in interesting, novel ways.

In addition to these benefits, there are several challenges mentioned in both applied and methodological studies. One of the major challenges is sampling. Twitter users are not a representative sample of any population, so results derived from Twitter data cannot generate conclusions about the general population (Gilbert, 2008). Because Twitter contains limited

socio-demographic metadata, it is hard to know which portion of the population is implicated in study results. Another issue is that traditional text-mining methods are not designed for such short pieces of texts, which can generate inconsistencies and irregular results (Hong & Davison, 2010). Because Twitter content is not verified or cleaned, datasets sometimes suffer from misspellings, data ambiguities, and inaccurate information. Finally, many studies indicate that the lack of context about why people tweet or what they are saying can make it difficult to interpret results.

3. Case Studies

3.1 Introduction to Case Studies

The case studies presented here draw upon the lessons from the literature overview to investigate which questions about cycling can be answered by analyzing Twitter data. Though it was not possible to avoid all issues associated with Twitter data, as that would be a very large task, we do take these difficulties into account in the design of our studies. We offer an in-depth reflection of these challenges in our discussion (Section 4), where we evaluate social media and Twitter analysis as a new tool for studying transportation mode choice.

Unlike traditional types of bicycling data, Twitter data offers access to unfiltered, direct statements that can be analyzed to learn about attitude and emotion. This type of data may elicit different insights than data delivered in response to a specific survey question. We offer a comparison of the advantages and disadvantages of these different methods in Table 3.

Table 3: Comparison of Twitter Analysis and other methods of cycling research

Research Method	Advantages	Disadvantages
Twitter data analysis	• Independently generated, unfiltered content • People use Twitter to express personal opinions and preferences • Large quantities of data can be collected quickly and cheaply	• Difficult to establish representativeness and accuracy of data • No socio-economic metadata
Traditional methods of cycling research (surveying and observation)	• Can ask direct questions • Can verify participant details in survey	• Survey format may prompt certain responses • Hard to collect large quantities of data • Potentially expensive

Twitter data is well suited to the use of an algorithmic toolkit, such as the methods discussed in Section 2.3.2, which can address different questions than more traditional methods of inquiry. A sample of these questions is included below. Previous social media studies show that Twitter can offer an interesting and dynamic view of human behaviors; the methods in these case studies seek to take advantage of the lexical complexity of the data by extracting meaning and finding patterns in the content. We seek to avoid some of the challenges of social media

analysis by framing our questions in terms of the community of content-generating users instead of the larger population, and preprocessing our data.

Potential research questions that can be answered using Twitter data:

- Why do people cycle?
- What barriers do bicyclists face?
- Is bicycling sentiment negative or positive in different cities?
- Which events, topics or incidents are most interesting to the bicycling community?
- What do people like about cycling? What do they dislike?
- For those who mention cycling but do not themselves cycle, why not?

3.2 Dataset

The dataset used in this study was composed of tweets from the NYC area gathered using the Twitter REST API. We collected NYC tweets using the Twitter API's geocode search parameter. We used the Twitter API's query parameter to collect tweets containing at least one of the following words: biking, bikes, bicycle, bicyclist, bike, cyclist, #biking, #bikes, #bike, #bicycle, #bicyclist, #cyclist. We collected approximately 1,100 tweets per day over a period of 3 weeks in March and April 2014, for a total of 18,401 tweets. We preprocessed the data by converting all words to lowercase, ignoring common words with fewer than three characters, and removing usernames and links. For example, applying the preprocessing steps to the tweet "RT @StreetsblogSF: .@WalkSF, @SFBike push SFMTA to make room for bike/ped projects in its budget http://t.co/Dqadhlu0hn" would yield "push sfmta make room for bike/ped projects its budget". Using the same process described above, we also collected cycling-related tweets from seven other U.S. cities with published bicycle rates. We collected approximately 20,000 tweets in each of these cities as well.

3.3 Sentiment Analysis

The research aim of the first case study was to understand the sentiment of bicycling-related content in U.S. cities, which is one of the proposed research questions that can help us evaluate the usefulness of Twitter analysis. We first performed sentiment analysis on our NYC dataset to ascertain whether the cycling tweets from NYC were positive or negative. Sentiment analysis is a well-documented and widely used natural language processing technique that assigns a sentiment score to pieces of text (Pang & Lee, 2008). This score can be binary (e.g., positive or negative) or take on multiple values (e.g., positive, neutral, or negative). We use it to gauge whether such methods can be reliably used on Twitter data and to learn about the sentiment of the tweets themselves.

Though widely implemented, there is no single way to perform sentiment analysis. With sentiment analysis, as with many machine-learning algorithms, one must build a classifier from a large body of documents that have already been correctly classified, called a corpus of training data. The algorithm looks at this pre-labeled training data to decide the correct class for new,

unlabeled data. In our study, we built a Naive Bayes classifier, because Naive Bayes has been successful in similar scenarios (Pak & Paroubek, 2010). With Naive Bayes, the probability of a given classification—e.g. positive or negative—is the likelihood of seeing a certain feature given a classification, multiplied by the prior probability of the classification. Mathematically, the probability of a classification, c, for a certain document, d, is given by: $P(c|x) \sim P(x|c)P(c)$, where x is a vector of feature values for each document d (Mitchell, 1997). The prior and likelihood values are calculated from the pre-labeled corpus of documents.

We used three different corpuses to train the classifier. The first corpus was a set of 498 hand-labeled tweets from Stanford University that had been classified as positive, neutral, or negative (Go et al., 2009). The second corpus, called the STS-Gold corpus, was a set of 2035 manually annotated tweets gathered by researchers at the Knowledge Media Institute in the UK, labeled as either positive or negative (Saif et al., 2013). Finally, we trained the classifier using a set of 25,000 labeled IMDB movie review texts from Stanford University, labeled as either positive or negative (Maas et al., 2011). For each corpus, we used a bag-of-words model, adopted from Luce (2012), which considered the presence or absence of each word in the tweet to be an individual feature. Each word in each tweet was compared against words from training data, and a label was selected based on the likelihood of seeing that combination of words. The results of our first classification attempts are shown in Table 4. Accuracy is defined as the total number of correct classifications divided by the total number of tweets.

Table 4: Results of first classification attempt

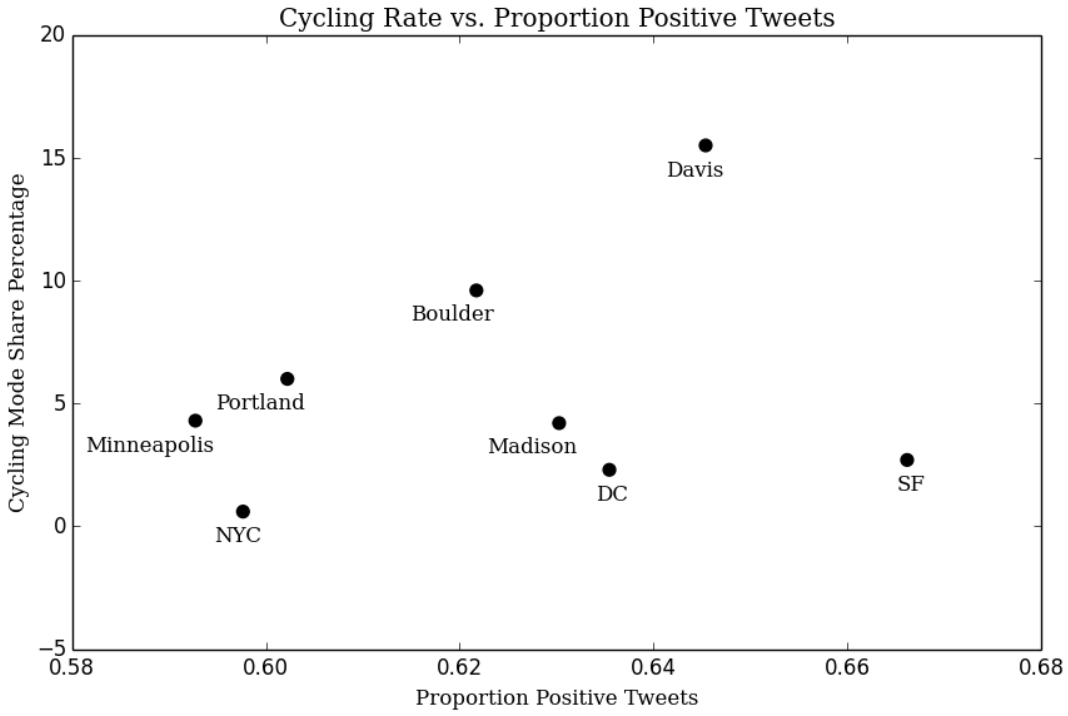
Classifier corpus	Accuracy on labeled test data	Percentage positive for NYC data
Stanford	0.83	0.59
STS-Gold	0.73	0.23
IMDB Movie Reviews	0.80	0.99

We found these results to be inconsistent; classifiers trained on different corpuses gave very different results. We thought that the results were perhaps skewed by the distribution of positive and negative tweets in the corpuses themselves. For example, only 9.8% of test tweets in the STS-Gold corpus are classified as positive, which perhaps accounts for the low percentage of positive NYC tweets (23%). To decide which classifier to use, we hand-labeled 500 randomly selected tweets from the NYC dataset to learn about the sentiment distribution in our NYC tweets. In this sample, 72% of tweets had positive sentiment, and 28% had negative sentiment. Using this hand-annotated dataset, we calculated the accuracy of each classifier, finding that the classifier based on the Stanford corpus produced the most accurate results for our dataset. We ran the classifier on our full data set of 18,401 tweets, finding that 59% of tweets in the NYC area were classified as positive.

We wanted more context to interpret this data point, so we looked at how the proportion of positive sentiment varied by city and region. Using the same process described above, we collected cycling-related tweets from seven other U.S. cities with published bicycle rates and

computed the proportion of tweets with positive sentiment about bicycling. These cities represent a wide geographical range and, like NYC, are similarly concerned with bicycling as a public issue. The correlation between the cycling rate and the proportion of positive tweets about bicycling is shown in Figure 1. The graph suggests some correspondence between Twitter data on bicycling and real-world metrics, which indicates the trustworthiness of our data.

Figure 1: Correlation between bicycling rates and cycling positivity in different US cities



3.4 Topic Modeling

The aim of the second case study was to discover the topics and events discussed within the NYC bicycling community by taking advantage of the lexical richness of the data. This aspect of Twitter data was particularly interesting because traditional data does not offer direct access to individuals' opinions. We used Latent Dirichlet Allocation (LDA), which, like sentiment analysis, is a widely used technique that has been applied to a variety of data types (Hong & Davison, 2010). LDA is an unsupervised machine-learning algorithm that discovers unknown topics in a series of sentences or texts by finding groupings of words that appear together frequently. We used it to uncover topics of discussion among Twitter-users mentioning bicycling. Using the `topicmodels` package in R, we ran LDA on a randomly chosen subset of 3,000 tweets from our NYC tweet dataset. A selection of the results, visualized in Table 5**,** show twelve topics and the ten words that contribute most to the meaning of each topic. Topics must be hand labeled; LDA output clusters of words that appear together often, but cannot understand them semantically. We looked at these word groupings and generated topic names based on their

contents. These topics offer an novel view into discussions of cycling, revealing some predictable topics like exercise and safety in addition to novel ones like coffee-cup holders and wall-mounted bike-storage that may be interesting to pursue further.

Table 5: LDA topics with words that contribute to each topic

Topic	Bike Share Memberships Prescribed	Central Park Rental Bikes	Cardio with Biking Commute	Bike Cup Holder	Citibike Debt	Effect of New Bike Lanes
Words that contribute to topic	doctors can prescribe collaboration overwhelmingly now share boston memberships patients	aggressive help park will central rental overly focus salespeople bicycle	time your out with ride cardio work distance speed mph	coffee holder recycled from made cans wood soda cyclists take	citibike investment star desperately was $14m share program needs nyc's	make life lanes better they how cyclists drivers affect certainly

Topic	New LED Bike Lights	Biking Gender Gap Closing	Solution to Citibike Debt	Suho Exo Promotes Bicycling	Wall-mounted Bike Storage	City Holds Seized Bikes
Words that contribute to topic	these safer keep lights road led bike-shaped never killed makes	but riders share nyc gap are 30-something dudes overwhelmingly narrowing	could schumer york bike save charles help fix struggling peddles	way healthy biking stress price-efficient relieve ride suho exo like	that easy mount makes wall decor extra better ever furniture	thousands release late bikes city including mathieu lefevre refuses seized

4. Discussion

4.1 Overview

These exploratory case studies are meant to illustrate the potential successes and challenges of social media analysis with Twitter data. They demonstrate that Twitter can offer some useful and interesting insights into cycling that would be hard to replicate with more traditional bicycling research. For example, using surveys, it would be difficult to gather enough qualitative data to identify the diverse and highly specific topics discovered using LDA. Twitter analysis can offer a richer, more dynamic view of the topics and concerns relevant to cyclists. While these results cannot speak for a population at large, they do seem to represent the topics of

interest to the population subset that is speaking about cycling. The results often center on particular incidents or ideas, such as Citibike debt, bicycle-friendly cup holders, or the seizure of NYC bicycles. These topics can shed light on how particular attitudes or concerns develop. While the proportion of bicyclists in the United States is still relatively small, learning about how ideas and trends develop in this community can help policymakers and the cycling industry adapt to the needs of both current and potential cyclists.

The case studies also reveal practical difficulties associated with social media analysis. While some of these challenges were anticipated because they appear in papers from the literature overview, others surfaced while carrying out these exploratory case studies. In the following sections, we discuss the challenges of Twitter data collection, sampling, analysis and interpretation, summarized in Table 6, offer some possible solutions, and end with some broader comments on the use of social media and Twitter analysis.

Table 6: Challenges of working with Twitter data, implications, and possible solutions

Category	Challenge	Implications	Possible Solutions
Data Collection	Hard to have fine-grained control over collection process	Imprecise results may misrepresent views of particular area	More sophisticated data pre-processing and filtering
	Poor location-tagging	Imprecise location results may miss relevant results or include irrelevant results	Exclusively collect geotagged data-points (generates much smaller dataset)
	Difficult to collect historical data	Results skewed by temporary trends	Use paid third-party service
	Hard to replicate data collection process	Replication important for scientifically rigorous process	Use tweet IDs to collect identical dataset
Data Sampling	Little or poor socio-demographic data	Difficult to establish representativeness of data	Infer socio-demographic information using other ML algorithms; use topic-based sampling and adjust research questions to account for lack of representativeness
	Unverified data	Difficult to establish truthfulness or accuracy of data	--
	Retweets or repeated data points	Repeated information; observations that are not independent	Eliminate retweets, only consider unique tweets; OR, incorporate retweets into research to track popular ideas
Data Analysis & Interpretation	Black-box techniques that only generate quantitative results	Misleading or incorrect results	Understand the implementation and tuning of different algorithms
	Overly general interpretation	Misleading or overly general results	Incorporate an understanding of social context

4.2 Data Collection

Data collection, a process often abstracted in papers, is an important part of social media analysis. As expected, it is easier to collect large amounts of Twitter data than traditional types of data. Whereas other methods require contacting participants, conducting interviews or dealing with equipment, Twitter data can be aggregated by simply running code from a computer. Using the search functionality from the Twitter REST API, we were able to collect thousands of data points from geographically diverse locations.

Nonetheless, Twitter faces its own set of challenges; we identify five main data collection challenges. The first is that it is hard to have fine-grained control over the data aggregation process. Because the data is in Twitter's hands and must be accessed using their API, data requests must use the parameters surfaced by the API. For example, in specifying the geocode parameter, one can only specify a latitude-longitude pair and a radius, which means that results come from a circular search area instead of the actual area of a city. If attempting to gauge the attitudes within a particular city, this imprecision may aggregate tweets that misrepresent or dilute the attitudes of that city. One solution may be to implement a more complex filtering system that post-processes tweets collected from Twitter, discarding those that do not fit within the city's actual boundaries.

Second, the way that Twitter processes location data can limit the utility of location-based results. A tweet is tagged with location if the user enables location-tagging and tweets from a location-enabled device, such as a mobile phone. If the user has not enabled location tagging, Twitter will use the "from" field from the user's profile if they have specified a particular town (Frank et al., 2013). Only 5-10% of the tweets collected were geo-tagged with a latitude-longitude pair. This method of location tagging means some relevant tweets will be omitted because they have no location, while some will be collected in error. One could avoid this issue by exclusively collecting tweets that are explicitly geotagged. The scope of the analysis in this paper did not require such fine-grained location data, but a study hoping to utilize location in a more analytical way might consider this option.

A third issue with Twitter API data collection is the difficulty of collecting historical data. The Twitter API only makes available data that is 1-2 weeks old, so it is more complicated to collect data about historical events or to collect data over a long period of time. The limited scope means that findings can be potentially skewed by temporary trends, such as weather or current events. While charting transitory trends may be the aim of some Twitter studies, this study hoped to develop more general findings. There are third-party companies affiliated with Twitter that can provide this service, but only at a cost.

Fourth, the data collection process is difficult to replicate. Though one could technically collect the same dataset by searching for tweets using their unique identifying numbers, running the same search command through the Twitter API will not gather an identical set of tweets, merely one that matches the same parameters. Another challenge is determining the proper amount of data to collect. While our dataset had 18,401 items, it is possible that a much larger data set would give rise to richer results. On the other hand, with too large a data set, one might encounter problems of overfitting and false topic discovery (Marcus & Davis, 2014). While

Twitter studies that use tweets to calculate a quantitative figure often use data sets with millions of tweets, the investigations in this paper were more qualitative in nature and were able to get interesting results without such massive datasets. Nonetheless, determining the proper amount of data is an open question, whose answer is often limited as much by capacity and availability as by conscious decision.

A final issue we encountered with this type of natural language data was data corruption or ambiguity, with spelling errors, slang, and multiple terms for the same concept. For example, in studying bicycling, we looked for tweets containing the words "bike", "bicycle" and "biking" (among others). However, words like "cycle" and "cycling" can refer to different ideas; other contextual information is required to understand if they refer to bicycling. We addressed this problem by only searching for words that explicitly referred to bicycling. While this solution allowed us to avoid irrelevant tweets, it also means that we probably missed some tweets that were indeed relevant.

4.3 Sampling

Tweets contain very little socio-demographic data, making it difficult to understand a sample in the context of the general population. While one might be able to infer such data from the provided metadata or user profiles, in general there is no information on age or gender. Additionally, it is difficult to assess the accuracy of the information in tweets because there is no requirement for truthfulness and no system in place to verify users' identities or the content of their tweets. The absence of verified data means conclusions from Twitter data are not representative of a larger population, which limits the power of these results. To address this issue, we adjusted our research to focus on ideas that develop entirely within the biking community, results that are potentially interesting to policymakers or cycling advocates.

Twitter data sampling is a challenge that has generated much discussion. According to Gilbert, sampling of Twitter data can either be classified as a probability technique, where researchers consider the sample representative of the entire population, or a non-probability technique, where the sample is not used to make inferences about the larger population (Gilbert, 2008). Because of the technical difficulties of ensuring access to a fully random sample, most studies, including the ones presented here, utilize a non-probability technique, such as topic-based sampling, which collects tweets that contain a specific hashtag or search query (Gerlitz & Rieder, 2013). As researchers, we consciously selected a non-representative subset of the full tweet population, which we used to gain insights about the behavior or feelings of that particular subgroup instead of the wider population.

Another Twitter-specific consideration is the phenomenon of repeated data points. Users can "retweet" a message, which means that they repost content that has been generated by another user. Popular tweets can be retweeted thousands of times. Researchers must decide how to deal with retweets in their analysis. On the one hand, if a tweet has been retweeted many times, it is a likely indication that the content of that tweet is popular and expresses the beliefs of many people. On the other hand, because it is a repeated data point, it does not contribute any

new content. Though we did not carry out any special treatment of retweeted data, it is a potentially interesting layer of analysis.

4.4 Analysis and Interpretation

As in any data analysis situation, it is important to blend human insights with quantitative results. Especially with machine learning, the human factor involved in the setup of algorithms, selection of features, and naming of topics can have huge influence over results. For example, we got very different sentiment analysis results when using different corpuses to train our classifier. This variation led us to hand-annotate a subset of our data to better understand which corpus was best suited for our data. Because humans decide the initial classifications that give rise to the classifiers, human opinions about whether something is negative or positive can have an outsize effect on the resultant classifications. When hand-annotating tweets, we encountered tweets that were ambiguous. For example, the tweet "I am going to start riding a bicycle. I am so sick of having to pay for car repairs" has both positive and negative sentiment. It makes a negative statement about the price of car repairs, but is optimistic about bicycles; we ultimately classified it positively. We found this process instructive in demonstrating how human insights can have an impact on quantitative processes.

These case studies demonstrate the importance of understanding the entire analysis process, not just the result. It is easy to abstract the analysis process with black box software methods. However, numerical results can suggest a misleading degree of certainty; thus, it is important to cross-validate results by testing different algorithms and setups, understanding the analysis process at a basic level, and not placing too much trust in any one value.

With social media data, it is particularly important to incorporate an understanding of social context when interpreting results. Especially in the U.S., where interest in cycling is a trending topic, we suspect that people taking the time to tweet about cycling hold strong views, either positive or negative. People may not tweet about their day-to-day habits, unless they face serious disruption, but by studying an aggregated mass of tweets, we may be able to learn about the habits of a group of people (Zook & Poorthuis, 2014). Alternately, studying these strong views may help us understand trends or developing attitudes, voiced by outspoken individuals, that will eventually be incorporated into a population's habits. Though understanding a group's Twitter use is complicated, it is important to interpret analysis results with careful consideration of the data's context.

4.5 Ideas for Further Research

These explorations confirm the feasibility of similar inquiries. With finer-grained location data, either from Twitter itself or a ride-tracking application like Strava or MapMyRide, one could link concerns or issues to specific areas of a city, which could help pinpoint smart infrastructure investments and alleviate hyper-local concerns. Another investigation could closely examine user interactions. Twitter is unique in offering access to these interactions. Studying them could yield insights about the cycling community in a particular city and allow

researchers to observe how users influence each other and spread ideas through a group of people. This understanding could help policymakers or cycling advocates spread information about the benefits of cycling. Finally, because users are identified by a unique number, Twitter analysis can follow a set of users throughout time—a longitudinal study—to see how their opinions and habits change over time and are molded by certain events.

It may also be productive to combine Twitter analysis with more conventional research methods. Twitter data could be used for exploratory research to discover topics and emotions that are mentioned frequently. In a second stage of research, traditional methods could be used to explore these topics more rigorously and in a more representative way. This method has some similarity to Grounded Theory in that it blends exploratory data analysis with more traditional techniques, and could this yield interesting results and help focus research efforts.

4.6 Overall Reflection

While acknowledging the potential of the Twitter case studies, this discussion also acknowledges the limitations of this data analysis method. Twitter research must be approached with an awareness of how human choices can impact results at any point in the analysis process. Researchers must carefully consider context when interpreting results. This methodology may not be valid in locations with limited Twitter use or where there is less cycling-focused Twitter activity. Studying how Twitter use varies between different cities and communities is out of the scope of this paper, but could enrich the results from this type of exploration. Additionally, an understanding of how Twitter use has changed in the eight years of its existence and will continue to evolve would also help guide data interpretation and development of research questions. These case studies demonstrate the limitations of Twitter analysis as a developing methodology, but also suggest it is a novel option that offers some compelling advantages over traditional research methods.

5. Conclusion

The aim of this paper is to evaluate social media analysis with Twitter data as a research method that can be used instead of or alongside more traditional cycling research methods. We conclude from our case studies that Twitter data—and social media data in general—offer some interesting advantages. Most importantly, the nature of the data allows researchers to address a new class of questions related to users' motivation, attitude and emotion. Twitter data benefits from unfiltered, direct insights from users, who aren't confined by the questions asked on a survey. We find that Twitter analysis can offer an accurate and dynamic view of the topics and concerns relevant to cyclists. We find that Twitter's lexically rich content differs significantly from the data generated by more traditional research methods and is suitable for complex, state-of-the-art machine learning techniques. Additionally, it is cheap and quick to collect a large volume of geographically diverse data. Twitter data offers the opportunity to build an understanding of the motivations and attitudes that underlie bicycle use. At the same time, we conclude from our case studies that there are a number of challenges specific to Twitter use,

which we used to launch a discussion of the benefits and possibilities of using Twitter data. We hope these insights will help increase the richness of bicycling research results in order to ultimately increase the feasibility and attractiveness of bicycling as a ubiquitous mode of urban transportation.

References

Agrawal, A. W., & Schimek, P. (2007). Extent and correlates of walking in the USA. *Transportation Research Part D: Transport and Environment*, *12*(8), pp. 548-563.

Bicycle and Pedestrian Provisions in (SAFETEA-LU) not Codified in Title 23. (2014, February 10). *Federal Highway Administration*. Retrieved March 6, 2014, from http://www.fhwa.dot.gov/environment/bicycle_pedestrian/legislation/legtealu.cfm#sec1807.

Conover, M. D., Ferrara, E., Menczer, F., Flammini, A., & Perc, M. (2013). The Digital Evolution of Occupy Wall Street. *PLoS ONE*, *8*(5), e64679.

Dill, J. (2009). Bicycling For Transportation And Health: The Role Of Infrastructure. *Journal of Public Health Policy*, *30*, pp. S95-S110.

Flegenheimer, M. (2013, December 29). Turning the City's Wheels in a New Direction. *The New York Times*. Retrieved March 6, 2014, from http://www.nytimes.com/2013/12/30/nyregion/turning-the-citys-wheels-in-a-new-direction.html.

Frank, M. R., Mitchell, L., Dodds, P. S., & Danforth, C. M. (2013). Happiness and the Patterns of Life: A Study of Geolocated Tweets. *PLoS ONE*, *8*(5), e64417.

Gerlitz, C., & Rieder, B. (2013). Mining One Percent of Twitter: Collections, Baselines, Sampling. *M/C Journal*, *16*(2).

Gilbert, G. N. (2008). *Researching social life* (3 ed.). London: Sage.

Go, A., Bhayani, R. & Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. *Processing*, pp. 1-6. Retrieved May 12, 2014 from http://www.stanford.edu/~alecmgo/papers/TwitterDistantSupervision09.pdf.

Heinen, E., Wee, B.V., & Maat, K. (2010). Commuting by Bicycle: An Overview of the Literature. *Transport Reviews*, *30*(1), pp. 59-96.

Heinen, E., & Handy, S. (2012). Similarities in Attitudes and Norms and the Effect on Bicycle Commuting: Evidence from the Bicycle Cities Davis and Delft. *International Journal of Sustainable Transportation*, *6*(5), pp. 257-281.

Hong, L. & Davison, B. D. (2010). Empirical study of topic modeling in Twitter. *Proceedings of the First Workshop on Social Media Analytics (SOMA '10), New York, NY*, July 25, 2010. Retrieved May 12, 2014 from http://dl.acm.org/ft_gateway.cfm?id=1964870&type=pdf&CFID=444091252&CFTOKEN=94294869.

Jacobsen, P. L. (2003). Safety in numbers: more walkers and bicyclists, safer walking and bicycling. *Injury Prevention*, *9*, pp. 205-209.

Krikorian, R. (2013, August 16). New Tweets per second record, and how!. *Engineering Blog*. Retrieved March 6, 2014, from https://blog.twitter.com/2013/new-tweets-per-second-record-and-how.

Krizek, K. J., Handy, S. L., & Forsyth, A. (2009). Explaining changes in walking and bicycling behavior: challenges for transportation research. *Environment and Planning B: Planning and Design*, 36, pp. 725-740.

Luce, L. (2012, January 2). Twitter sentiment analysis using Python and NLTK. *Laurent Luce's Blog*. Retrieved May 12, 2014, from http://www.laurentluce.com/posts/twitter-sentiment-analysis-using-python-and-nltk/.

Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y. & Potts, C. (2011) Learning word vectors for sentiment analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 1, Portland, Oregon*, June 2011. Retrieved May 12, 2014 from https://www.aclweb.org/anthology-new/P/P11/P11-1015.pdf.

Marcus, G., & Davis, E. (2014, April 6). Eight (No, Nine!) Problems With Big Data. *The New York Times*. Retrieved May 12, 2014, from http://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html?nl=todaysheadlines&emc=edit_th_20140407.

Martínez-Cámara, E., Martín-Valdivia, M. T., Ureña-López, L. A., & Montejo-Ráez, A. R. (2014). Sentiment analysis in Twitter. *Natural Language Engineering*, *20*(1), pp. 1-28.

Mitchell, T. M. (1997). Bayesian Learning. *Machine Learning* (pp. 154-199). New York: McGraw-Hill.

Pak, A. & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta*, May 2010. Retrieved May 12, 2014 from http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf.

Panagiotopoulos, P., Bigdeli, A. Z., & Sams, S. (2012). "5 Days in August" – How London Local Authorities Used Twitter during the 2011 Riots. *Electronic government*, pp. 102-113. Berlin: Springer.

Pang, B., & Lee, L. (2008). Opinion Mining And Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), pp. 1-135.

Pucher, J., & Buehler, R. (2008). Making Cycling Irresistible: Lessons from The Netherlands, Denmark and Germany. *Transport Reviews*, *28*(4), pp. 495-528.

Pucher, J. R., & Buehler, R. (2012). International Overview. *City Cycling* (pp. 9-29). Cambridge, Mass.: MIT Press.

Saif, H., Fernandez, M., He, Y. & Alani, H. (2013). Evaluation Datasets for Twitter Sentiment Analysis: A survey and a new dataset, the STS-Gold. *Proceedings of the First ESSEM workshop*, 2013. Retrieved May 12, 2014 from http://ceur-ws.org/Vol-1096/paper1.pdf.

Schoonderwoerd, N. (2013, November 7). 4 ways how Twitter can keep growing. *PeerReach Blog*. Retrieved March 5, 2014, from http://blog.peerreach.com/2013/11/4-ways-how-twitter-can-keep-growing/.

Shaheen, S. A., Guzman, S., & Zhang, H. (2010). Bikesharing in Europe, the Americas, and Asia: Past, Present, and Future. *Transportation Research Record, 2143*, pp. 159-167.

Trimm, H. (2012, October 11) New bike counter logs bridge trips. *Seattletimes.com*. Retrieved May 9, 2014 from http://seattletimes.com/html/picturethis/2019409280_cycle12.html.

Troiano, R. (2005) A timely meeting: objective measurement of physical activity. *Medicine and Science in Sports and Exercise,* 37, pp. S487 - S489.

Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. Menlo Park, CA: The AAAI Press, pp. 178-185.

twitter.com. (n.d.). *Alexa: The Web Information Company*. Retrieved March 6, 2014, from http://www.alexa.com/siteinfo/twitter.com.

Zook, M., & Poorthuis, A. (2014). Offline Brews and Online Views: Exploring the Geography of Beer Tweets. *The Geography of Beer*, eds. M. Patterson and N. Hoalst-Pullen. Springer. pp. 201-209.