Exploitation of Tweets to Measure Experienced Utility

Allison Madigan
University of Colorado Boulder

Follow this and additional works at: http://scholar.colorado.edu/honr_theses

Recommended Citation
Exploitation of Tweets to Measure Experienced Utility

By

Allison Madigan

International Affairs Departmental Honors Thesis, University of Colorado Boulder

Thesis Advisor:
Carlos Martins-Filho, Department of Economics

Defense Committee:
Vicki Hunter, Department of International Affairs
Robert McNown, Director of International Affairs Program, Department of Economics
# Table of Contents

Abstract ................................................................................................................................. 3

Introduction ............................................................................................................................ 4

Literature Review .................................................................................................................... 7
  Understanding the Distinction between Experienced Utility and Decision Utility ........... 7
  Experience Sampling; Scaled up, by Twitter ................................................................. 8
  Twitter ................................................................................................................................. 10
  Demographics of Twitter Users ...................................................................................... 11
  Emergence of Twitter Data in Academia ......................................................................... 13
  Valuation of Public Goods and Environmental Externalities ....................................... 17
    Stated Preference Methods .......................................................................................... 18
    Revealed Preference Methods ................................................................................... 19
    Life Satisfaction Approach ....................................................................................... 19

Methodology ......................................................................................................................... 21
  Tweet Collection ............................................................................................................. 21
  Sentiment Analysis of Tweets ......................................................................................... 23
    Sentiment Lexicons ..................................................................................................... 24
  Data Set .............................................................................................................................. 25
  Tweet Analysis Overview ............................................................................................... 27

Results ................................................................................................................................... 29
  Data integrity ................................................................................................................... 29
  Regression Analyses ....................................................................................................... 34
    Exploratory Hypotheses Based on A Priori Predictions ............................................ 34

Discussion ............................................................................................................................. 40

Caveats and Future Directions ............................................................................................ 46
  Statistical Caveats .......................................................................................................... 46
  Data Set Combination .................................................................................................... 46
  Hardware and Time Limitations .................................................................................... 48
  Potential for Sentiment Analysis; Machine Learning Algorithms .............................. 48
  International Perspective ............................................................................................... 50
  Collaboration ................................................................................................................... 51

Conclusion ............................................................................................................................ 52

Postface ................................................................................................................................. Error! Bookmark not defined.

References ............................................................................................................................ 54
Sentiment analysis of tweets

**ABSTRACT**

The classic microeconomic utility model makes stringent assumptions about the preferences of individuals; that preferences are well defined and consistent and that individuals are rational actors who choose to maximize their preferences. However, evidence from psychology and behavioral economics suggest those assumptions do not reflect a realistic understanding of human behavior. Instead, a less strict and more accurate interpretation of utility might be the hedonic experience associated with different outcomes, otherwise known as *experienced utility*. Research is ongoing on how to measure experienced utility. The prevailing metric for experienced utility is subjective well-being survey responses, however those have their shortcomings. The current study investigates and tests the feasibility of using large-scale sentiment analysis of tweets to arrive at an experienced utility metric for cities.
INTRODUCTION

Societies are interested in themselves. We want to know how we’re doing. Are we prosperous? How do we stack up next to others? In 1974 Richard Easterlin asked *Does Economic Growth Improve the Human Lot?* and found that despite steady increases in per capita income in the US between 1946 and 1970, responses to subjective well-being surveys remained fairly constant (Easterlin, 1974). This seminal paper, now referred to as the Easterlin Paradox, called into question the parameters against which, as a society, we have been evaluating ourselves. GDP, per capita income and economic growth represent a few of the go-to measures countries turn to, to answer the question, *How do we stack up next to other societies?* The metrics by which societies assess their prosperity have begun transitioning away from being exclusively monetary based. The United Nation Millennium Development Goals included: universal education, gender equality, child health, maternal health and environmental sustainability among others.

What prompted for so long, our unwavering faith in monetary metrics to evaluate our societal well-being? Arguably, the prevailing theory of utility in economics may be responsible. This theory assumes individuals are rational actors who act on complete and consistent preferences to maximize utility. Put differently, “individuals are rational-decision makers who know what they want and how to get it, or to get as close as possible given their budget” (Helliwell, Layard, & Sachs, 2012 p. 5). Despite the prevalence of this popular notion, some are skeptical. Psychologist and Nobel Laureate in Economics, Daniel Kahneman, has spent the better part of his career debunking the assumptions of classic microeconomic utility theory in favor of an alternative interpretation of utility, he refers to as *experienced utility*. Experienced utility is the hedonic experience, or moment-to-moment experiences of pleasure and/or pain, associated with an outcome (Kahneman, Wakker, & Sarin, 1997). One fundamental difficulty and contributor to many economists’ hesitation to jump on board, is that it is difficult to observe and/or quantify an individual’s subjective hedonic experience. Fortunately, the tides are changing.

The United Nations has recently released the first ever *World Happiness Report*. The report focuses on the importance of subjective well-being surveys for understanding what factors cause some societies to be happier than others. They view subjective well-being survey responses as measures of experienced utility. They’re not the first to explore this possibility, in fact, it’s a central premise of a landmark study of the
Sentiment analysis of tweets

valuation of public goods and costs of public bads; *The Life Satisfaction Approach* (Frey, Luechinger, & Stutzer, 2004). The life satisfaction approach models subjective well-being survey responses as functions of income, quantity of the public good being valued, and vector of socio-demographic and socio-economic control variables.

Subjective well-being surveys represent one possible option for measuring experienced utility but not necessarily the best or most efficient. I explore and test the feasibility of using tweets from the micro blogging service, Twitter, to remotely monitor the experienced utility of cities. In the current study, sentiment of Tweets will be aggregated by city to get a general idea of the city’s positive or negative-ness. Twitter is an online micro blogging web service that allows users to ‘tweet’ brief (140 characters or less) status updates pertaining to any topic and as often as they would like. In the current study, a large sample of Tweets from 366 metropolitan statistical areas is collected. Simple text analysis is applied to them to gauge positive and negative affect. This method is similar to *experience sampling*, a method used in the field of psychology/psychoinformatics that is designed to gauge an individual’s emotional state. However, the current study attempts experience sampling on a much wider scale by utilizing Twitter. Experience Sampling is considered to be, by D. Kahneman, one of the best methods for measuring experienced utility, but, in its traditional pen-and-paper form, it is severely limited by cost and scope (Kahneman & Krueger, 2006). Twitter data is, scalable to large populations, abundant, essentially free, and ruthlessly current. All of these points address the weaknesses of traditional subjective well-being surveys.

The life satisfaction approach has laid the groundwork for the cost analysis of public goods and public bads using experienced utility. The current study investigates the adaptation of the life satisfaction approach to predict sentiment analysis of tweets rather than subjective well-being survey responses. We hypothesize that affect present in tweets aggregated by city is a valid measure of experienced utility of cities. The current study is also a demonstration that a single researcher with entry level programming experience and a personal laptop has the capability to collect and analyze massive amounts of data and infer, through text sentiment analysis, a useful surrogate measure of experienced utility.
This paper is organized into the following chapters: literature review, methodology, results, discussion, caveats and future directions, and conclusion. The literature review chapter consists of the subsections: understanding the distinction between experienced utility and decision utility, experience sampling; scaled up, by Twitter, Twitter, demographics of Twitter users, emergence of Twitter data in academia, valuation of public goods and environmental externalities, stated preference methods, revealed preference methods and the life satisfaction approach. The methodology chapter consists of the subsections: tweet collection, sentiment analysis overview, sentiment lexicon, tweet processing, collection of independent variables and function set up. The results chapter presents the results of the analysis. The discussion chapter discusses the interpretation of significant findings and the capacity of Twitter as a data collection tool. The future directions chapter discusses limitations of the current study and how to improvement tweet collection and sentiment analysis in the future. Additionally, the future directions chapter discusses the international relevance of Twitter as a data collection tool and the potential for future applications in an international sense. Finally, the conclusion chapter reiterates the salient themes of the study.
LITERATURE REVIEW

UNDERSTANDING THE DISTINCTION BETWEEN EXPERIENCED UTILITY AND DECISION UTILITY

The classic utility model has been acknowledged to be inadequate in various scenarios, because it makes overly stringent assumptions about the rationality and preferences of individuals. One domain in which it’s shortcomings are vividly evident is that of valuing public goods and assessing the costs imparted by negative externalities. This is best illustrated by an example. The economist or social planner might like to know the welfare benefit to a community of a proposed network of bike paths around the city akin to the one in Boulder, Colorado. Under the classic utility theory the scope of the welfare gain is represented by the aggregate of individuals in the communities’ willingness to pay (WTP) for this network of bike paths. To ascertain individuals’ WTP, economists often rely on contingent valuation surveys, which might ask, “How much would you be willing to pay for the installment of a bike path network around the city?” or “What hypothetically would it be worth to you if a network of bike paths was added to the city?”

This approach assumes: 1. The individual is privy to all possible benefits and costs of the bicycling network, or possesses “complete information” 2. The individual evaluates this information with respect to their budget. If a person who had previously taken advantage of such a bike path system had been asked, they might have a more complete understanding of the benefits associated with the creation of bike network, and estimate a higher willingness to pay. Alternatively, imagine asking a respondent who didn’t grow up in a biking community, was never genuinely introduced to bicycling, or who had never even considered bike commuting as an alternative form of transportation. Would one expect that respondent to attribute much value to the creation of a bike path network? We cannot reasonably expect the second respondent or even the general public to consider the expansive repercussions of introducing a bicycle path network such as: improved health and reduced obesity in the local population, reduction in car accidents, reduction in drunk driving incidents and reduction in carbon dioxide emissions, to name a few. This is especially true given the time, energy, and resources that have been devoted by academics, non-profit organizations and government to identify these...
Sentiment analysis of tweets

ripple effects. And given the similarity between this issue example and many other public goods debates, we can conclude that individuals are not privy to the benefits/costs of public goods questions.

This example illustrates a chief motivation for the life satisfaction approach; it is less cognitively demanding because individuals are not required to value hypothetical goods. Furthermore they aren’t tasked with identifying the complex interdependence between their level of well-being and factors in their lives which contribute to that, such as living somewhere with a complex bikeway system.

EXPERIENCE SAMPLING: SCALED UP BY TWITTER

One of the principal criticisms to experienced utility approaches is that a person’s subjective hedonic experience cannot be objectively measured. So far, researchers have relied heavily on subjective well-being survey responses to estimate experienced utility. Tella et al. (2007) investigate the tradeoffs between inflation and unemployment by linking them to happiness survey responses. Similarly, Levinson (2012) and MacKerron & Mourato (2009) use subjective well-being survey responses to value air quality. Praag & Baarsma (2001) explore the shadow costs of airport noise nuisance, incorporating both subjective well-being survey responses and the hedonic market approach in their investigation.

Despite their popularity, subjective well-being surveys may not represent the optimal choice. One problem with them is their susceptibility to variations in the circumstances during which the questionnaire is submitted. Questions are typically designed to elicit global evaluations about how satisfied the respondent is with their life (e.g. - All things considered, how satisfied are you with your life as whole these days? from the World Values Survey); answers are easily swayed by the most recent experiences of the respondent. For example, consider the classic experiment by Schwarz (1987) discussed in (Kahneman & Krueger, 2006 p. 6):

Schwarz (1987) invited subjects to the lab to fill out a questionnaire on life satisfaction. Before they answered the questionnaire, however, he asked them to photocopy a sheet of paper for him. A dime was placed on the copy machine for a randomly chosen half of the sample. Reported satisfaction with life was raised substantially by the discovery of the coin on the copy machine—clearly not an income effect.
Sentiment analysis of tweets

Nevertheless, a number of studies have also produced findings, which validated responses to subjective well-being surveys. Kahneman and Krueger (2006) surveyed several studies which found correlations between reported life satisfaction and other indicators of happiness including smiling frequency, ratings of one’s happiness by friends, frequent verbal expressions of positive emotions, sleep quality, happiness of close relatives, and self-reported health.

Subjective well-being surveys are also costly and time consuming. Six waves of World Values Surveys have been conducted since 1981. The World Values Survey has acquired data from 102 countries at least once since 1981 (http://www.wvsevsdb.com/wvs/WVISintegratedEVSWVS.jsp?Idioma=I). Each wave took roughly four years and not all participating countries were able to be surveyed in each wave. Twitter data, on the other hand, is free, with marginal increases in costs associated with increases in the size of the data collection. Additionally, in stark contrast to a minimum of four years between surveys, Twitter is available in real time. The question of whether Twitter is a plausible measure for experienced utility still awaits discussion.

Another technique for acquiring a measure of experienced utility is experience sampling (Kahneman & Sugden, 2005), which is an approach created to collect information on individuals’ described feelings in real time (Csikszentmihalyi, 1990). In this approach, participants are asked to carry a hand held computer device with them throughout their normal daily routine. During each day of the experiment, participants are prompted several times by the device to answer questions about their current activity, physical location, who they are with, their current feelings, and the intensity of those feelings. Researchers assess participants experiences throughout the day looking at the degree of positive affect (the range of positive emotions) and negative affect (the range of negative emotions) exhibited by the participant through their activities and descriptions of their feelings. Episodes of positive affect and negative affect are generally measured by their self-reported frequency and intensity. Frequency of pleasure moments is a measure of positive affect while pain experiences are a measure of negative affect. As such, experience sampling appears extremely in tune with experienced utility. The difficulty arises in scalability. While this method promises to be very accurate, the technology resources and time required of the participant severely limits sample size. While some of this difficulty might be mediated by advances in handheld computer technology and their expanding presence in societies (i.e. - smart
Sentiment analysis of tweets

phones), there remains the challenge of convincing and making it worthwhile for participants to dedicate a significant amount of time throughout the day to answering questions.

Experience sampling, as described above, is essentially a means for monitoring individuals’ fluctuations in positive affect and negative affect throughout the day. In the case of the current experiment, the conversational nature of Twitter promotes this kind of frequent status updates. Curiously, participants in experience sampling studies receive monetary compensation for their cooperation, while Twitter users voluntarily ‘tweet’ brief status updates throughout the day. Golder and Macy (2011) actually follow individual Twitter accounts and monitor shifts in positive and negative affect with respect to the time of day. However, the current study suggests the aggregation of tweets, based on their city of origin, can allow measurement of experienced utility of whole metropolitan areas (cities/towns).

A big step being taken in the current study is the application of theoretical concepts developed with regard to individuals to measure whole groups of individuals. Twitter data is one of many types of ‘big data’ that has begun to be used for analysis of group and trends. Google trends is another example of a big data set that has been utilized to make fairly accurate predictions about ongoing trends and interests. For example, Google trends monitors and predicts flu activity around the world by tracking where people are searching for topics related to the flu (http://www.google.org/flutrends/).

**Twitter**

Twitter is an online social networking and micro-blogging website through which users share short (140 characters or less) texts. The texts, referred to as “tweets” contain a medley of information, from, simple updates on users’ status or current activity, to news opinions, current events, photos, or videos. Unlike other social networking services such as Facebook, where users carefully monitor their privacy settings, restricting full profile visibility to people whom they’ve authorized to do so or “accepted as friends”, Twitter exists in a particularly public domain. Approximately 91% of Twitter users choose to make their profile and tweet history publicly accessible and visible (Mislove et al.). Twitter is a space where people voluntarily shout out to the rest of the world their thoughts, feelings, opinions and essentially anything else they find interesting. From the
Sentiment analysis of tweets

point of view of the researcher, the public nature of Twitter allows the bypassing of usual rules for obtaining consent from participants.

Since officially launching on July 15th, 2006, Twitter has experienced huge growth in accounts and daily tweets. Within two years one billion tweets were tweeted; one year later the five billion tweets mark was passed. By 2010 over 50 million tweets were appearing daily. (http://mashable.com/2011/05/05/history-of-twitter/ accessed 2/4/13) Near the end of 2012, there were an estimated seventy-two million active twitter accounts with each account averaging five tweets a day. Three hundred fifty-seven million accounts have posted at least once, and ninety-six million have tweeted at least once in the last thirty days. (http://mashable.com/2012/12/11/twitter-1-billionth-user-id/ accessed 2/4/2013) Finally, Twitter is used in nearly every country around the world and supports over twenty languages. (https://twitter.com/about accessed 2/4/13)

**DEMOGRAPHICS OF TWITTER USERS**

Given Twitter’s widespread use and accessibility to researchers, tweets represent an increasingly intriguing data set. However, before diving in, important considerations of the demographic makeup of twitter users are in line. Few have tackled this question since little information regarding the demographics and the socio-economic status of the Twitter-using population is made explicitly available. Mislove et al. explores the geographic distribution, gender and race/ethnicity of the Twitter using population. They deduce that Twitter users over represent populous counties by observing that as populations of counties increase, the *Twitter representation rate* increases (number of twitter users in a county divided by number people in the county in the 2000 US census). The result is oversampling of more populous counties, however, they find no other geographical bias. That is, all highly populated cities from the West Coast, to the East Coast, to the Midwest and to the South follow the same general pattern. The study also investigates the gender of Twitter users. They do so by cross-referencing the self-reported first names associated with each user’s profile with lists of female and male names. These lists are compiled from the lists of the 1000 most popular male and female names for babies born in the U.S. for each year 1900-2009 reported by the U.S. Social Security
Sentiment analysis of tweets

Administration.¹ Through this method they found a stronger male presence; 71.8% of name matches corresponded to male names. Finally the study aimed to explore the race/ethnicity of twitter users. The US census releases the distribution of race/ethnicity for high-frequency (over 100 census respondents) last names. Seventy-one point eight percent of self-reported Twitter user last names matched with the 2000 U.S. census last name race/ethnicity information. From here a cross county comparison of race/ethnicity and twitter user race/ethnicity was conducted. The identified trends included under sampling of Hispanics in the Southwest, under-sampling of African-American users in the South and Mid-west, and over-sampling of Caucasian users in major cities. The Twitter data sample for this analysis was accumulated between March 2006 and August 2009. Certainly, given the discussed growth of Twitter since then, there is reason to imagine the demographics of Twitter users might have evolved.

More recently, the Pew Research Center, a subsidiary of The Pew Charitable Trust, conducted research regarding the demographics of the Twitter population in survey form. The report, *Twitter Use 2012*, was formulated as part of the Pew Internet and American Life Project. This demographic information is different than that previously presented, in that it was collected through phone surveys and subcategorized by Internet users.² That is, the numbers are present in the form “% of Internet users with each group who use Twitter”. Interesting findings include: 15% of Internet users use Twitter, including 8% of Internet users doing so in a typical day. Table 1 presents some of the information gathered in the Pew report. There doesn’t appear to be different usage patterns between males and females but there is a skew toward young adults. Also notable is that a much larger proportion of African American internet users use Twitter than whites or Hispanics. The results about annual housing income are also interesting because there appears to be no skew

---

¹ 241 names, which occurred in both lists, were removed (gender ambiguous). They found name matches for 64.2% of the users.

² Princeton Survey Research Associates International conducted phone interviews for the Pew Research Center’s Internet & American Life Project in English and Spanish between January 20 and February 19, 2012. 33,732 landline and 22,499 cellphone, numbers were dialed from random digit dial (RDD) generation; 1,352 landline surveys were completed and 901 cellphone. Internet users described by the report landed at n=1,729.
Sentiment analysis of tweets

towards wealthier brackets. Geographic location is also predictive of twitter usage with Internet users living in urban or suburban areas using Twitter more than those in rural parts of the country.

Table 1. Twitter demographics

<table>
<thead>
<tr>
<th>Demographic</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>All adult Internet users</td>
<td>1729</td>
<td>15</td>
</tr>
<tr>
<td>Men</td>
<td>804</td>
<td>14</td>
</tr>
<tr>
<td>Women</td>
<td>925</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>18-29</td>
<td>316</td>
<td>26</td>
</tr>
<tr>
<td>30-49</td>
<td>532</td>
<td>14</td>
</tr>
<tr>
<td>50-64</td>
<td>521</td>
<td>9</td>
</tr>
<tr>
<td>65+</td>
<td>320</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race/ethnicity</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>White, Non-Hispanic</td>
<td>1229</td>
<td>12</td>
</tr>
<tr>
<td>Black, Non-Hispanic</td>
<td>172</td>
<td>28</td>
</tr>
<tr>
<td>Hispanic</td>
<td>184</td>
<td>14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Annual household income</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than $30,000/yr</td>
<td>390</td>
<td>19</td>
</tr>
<tr>
<td>$30,000-$49,999</td>
<td>290</td>
<td>12</td>
</tr>
<tr>
<td>$50,000-$74,999</td>
<td>250</td>
<td>14</td>
</tr>
<tr>
<td>$75,000+</td>
<td>523</td>
<td>17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Education level</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No high school diploma</td>
<td>208</td>
<td>22</td>
</tr>
<tr>
<td>High school grad</td>
<td>465</td>
<td>12</td>
</tr>
<tr>
<td>Some College</td>
<td>447</td>
<td>14</td>
</tr>
<tr>
<td>College+</td>
<td>698</td>
<td>17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Geographic location</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban</td>
<td>520</td>
<td>19</td>
</tr>
<tr>
<td>Suburban</td>
<td>842</td>
<td>14</td>
</tr>
<tr>
<td>Rural</td>
<td>280</td>
<td>8</td>
</tr>
</tbody>
</table>

Source: Pew Research Center’s Internet & American Life Project Winter 2012 Tracking Survey

Emergence of Twitter Data in Academia

Twitter users may not completely represent the general population. Nevertheless, tweet discourse analysis is tepidly entering the academic realm. A number of recent studies have utilized tweets. These studies have largely explored the forecasting potential of tweets including predicting movie box office sales, stock market movements, and elections (Asur & Huberman, 2010)(Bollen, Mao, & Zeng, 2011)(Tumasjan & Sprenger, 2010). Predictions of movie box office sales rested on the hypothesis that the more chatter surrounding a movie, the more likely people would go see it. In fact, the model based on Twitter chatter
developed in the study surpassed the Hollywood Stock Exchange in accuracy (Asur & Huberman, 2010). Theorizing that public mood and sentiment may play an influential role with regard to stock market values, Bollen, Mao and Zeng turned to Twitter to measure public mood and sentiment. They used an algorithm called GPOMS to classify tweets along 6 mood dimensions and found that some were Granger causative of the DJIA (Dow Jones Industrial Average). Twitter is also being employed to detect and monitor disease outbreaks (Culotta, 2010)(Lampos, Bie, & Cristianini, 2010). In a similar vein, one study investigated Twitter’s capacity to detect real-time events, such as earthquakes (Sakaki, 2010).

The promise of Twitter to probe topics and predict elections in the political arena arouses interest on both academic and commercial fronts. The “Twitter Political Index” is a service that kept track of mentions of the then presidential candidates, Obama and Romney, and analyzed the sentiment of the tweet in which they were mentioned. The Twitter Political index is powered by TOPSY a “real time social analytics provider”, currently it offers information about a variety of political topics such as abortion, the economy, education, immigration etc. (https://election.twitter.com/#topics/abortion)(http://about.topsy.com/election/). Within the domain of Academic, tweet analysis has been used to predict voting patterns in two ways: monitoring reactions to events such as political debates, predicting elections by estimating popularity (Diakopoulos & Shamma, 2010)(Tumasjan et al. 2010).

Quite a few academic papers have also sought to predict elections through collection and analysis of tweets (Bermingham & Smeaton, 2011)(Tumasjan et al., 2010)(Chung & Mustafaraj, 2010)(O’Connor, 2010)(Tumasjan & Sprenger, 2010). Sheer frequency of mentions (without the use of sentiment analysis) of the candidates or parties in the election in tweets produces highly predictive results with margins of error in the ranks with traditional survey polling methods. Bermingham and Smeaton, while investigating the predictive power of various measures in the 2011 Irish General Election, found that of non-parametric methods, volume of tweets mentioning the different parties was the most predictive of election results but that the proportion of party mentions being in a positive context out of all positive tweets came in a close second (Bermingham & Smeaton, 2011).
Sentiment analysis of tweets

As a preliminary condition for the validity of predicting elections with Twitter data, some studies have endeavored to establish Twitter as a space for political deliberation. For example Tumasjan et al. seeks to confirm Twitter as a forum for political debate by examining how it lines up with the commonly accepted criterion, “the exchange of substantive issues and the equality of participation” (Tumasjan et al., 2010). To do this they randomly select a small portion of tweets associated with each political candidate and manually assess their substance. To establish exchange, they find the proportion of their collected tweets which contain the “@” sign. Under the Twitter format, including an “@” followed by the user id of another person, the tweet will be sent to that person. Consequently “@” signs are often present within conversations between Twitter users. This measure of addressivity has been formally investigated and confirmed by (Honeycutt & Herring, 2009). Through their evaluation they conclude that Twitter is a place substantive discussion rather than just a place for people to post their own opinions but that it is dominated small proportion of users.

The question of whether Twitter is forum for political deliberation draws parallels in the investigation of Twitter as tool for measuring public mood. In order to use Twitter to monitor public mood, one must consider whether Twitter can be considered a forum where people express their mood. No criteria as concrete as the frequency of “@” symbols, exists for determining whether mood is being expressed. The primary reason for this is double sided, namely a person does not need to be speaking specifically about their emotional state for one to gauge their mood. In contrast, positive and negative affect is often considered indicative of mood in psychology. Extracting positive affect and negative affect is achieved through a variety of methodologies in psychology. One strategy for establishing Twitter as a space where mood is expressed is to determine whether Twitter is sufficiently similar to previously established measurements. Compared to literature surrounding election prediction with Twitter, those using Twitter as an indicator of public mood often overlook this important step. After surveying the different methodologies for abstracting positive and negative affect, this paper argues that Twitter can be likened to Experience Sampling but at the city level rather than the individual level.

Some inquiry into employing Twitter to track public mood and/or measure happiness similar to the aim of this paper has begun. Distinct from most Twitter studies, Golder and Macy (2011) collect tweet histories of individuals. They use the Linguistic Inquiry and Word Count (LIWC) lexicon to apply a text
Sentiment analysis of tweets

analysis to determine the positive affect and negative affect of tweets. For each individual they track fluctuations in positive affect and negative affect through the day and week, with the goal of modeling diurnal and seasonal mood rhythms. The motivation of the study is to better understand mood patterns because “individual mood is an affective state that is important for physical and emotional well-being, working memory, creativity, decision-making, and immune response” (Golder & Macy, 2011 p. 1878). In contrast, (Dodds et al. 2011) seeks to gauge, characterize, and understand the well-being of large populations by analyzing tweets. They develop their own lexicon based solely on frequency of word usage, which includes happiness scores for over 10,000 individual words. They employ this word set to remotely sense fluctuations in happiness across various timescales including hours, days, months, and years. They refer to their methodology as a tunable “hedometer”. (Bollen, Pepe, & Mao, 2011) support the idea that Twitter can be utilized to ascertain a measure of public mood by demonstrating that movements in aggregated Tweet sentiment can be linked to notable socio-economic events, such as the 2008 election results, Thanksgiving, the failure of several international banks, and shifts in the price of crude oil. In an unpublished study, Pulse of the Nation: US Mood Throughout the Day inferred from Twitter, researchers from Northeastern University and Harvard University, apply similar lexicon based sentiment analysis techniques to the above studies to infer happiness levels, and then use cartograms and time lapse videos to disseminate their findings in mood variation over time and with regard to geographic location and density of tweets originating from different regions (A Mislove, Lehmann, & Ahn, 2010).

Two other studies aggregate tweets in cities. Hannak, Anderson and Barrett (2012) in their study Tweetin’in the Rain: Exploring the societal-scale effects of weather on mood employ advanced, machine learning techniques to model mood in tweets and then predict said mood with variation in weather. Variation in mood and weather is achieved through cross sectioning by cities in the US. Finally, in The Geography of Happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place In (Mitchell et al., 2013) happiness is determined from sentiment analysis of tweets and mapped to geographic places around the US, including states and cities. They are concerned with determining what about a geographic place factors into societal levels of happiness.
Valuation of Public Goods and Environmental Externalities

One of the primary goals of environmental economics is the valuation of public goods and environmental externalities. While public goods and environmental externalities or public bads affect the overall well-being of a society, their value is difficult to capture and they are often ignored by traditional markets. This is called a market failure and is often the economic justification for government provision and intervention. Arthur Pigou was one the first scholars in 1920 to argue for the internalization into the market of negative externalities resulting from pollution by the requirement that the polluters pay a tax equal to the marginal social cost of polluting emissions (Nimubona, Sinclair-desgagné, & Paris, 2005). As in the case for the Pigouvian tax, government is in a unique position to facilitate social cooperation in areas where individual incentive is lacking (Hardin, 1968).

Hence, identifying the marginal social cost or benefit of a public good or public bad is of upmost importance, particularly when the government or social entity must choose where to direct its limited resources. However, the qualities that define public goods and environmental externalities inherently create challenges to their valuation, especially subject to the classic utility theory framework (B. Frey et al., 2004). Two distinct components characterize a public good: non-rival consumption and non-excludability (Buchanan, Musgrave, 1999). Again, a bike path example helps exemplify these two concepts: Sally’s decision to ride her bike along a new safe bike path to work on Monday has no effect on Joe’s choice to ride his bike to the gym Monday morning. This is known as non-rival consumption. Likewise, Joe can’t stop Sally from also riding on the bike path, which is called non-excludability. Consider perhaps that Joe knows Sally is extremely determined to construct a bike path network, and that once the bike path is built, he can take advantage of the bike path as much as he’d like, regardless of whether he contributes to its production or not. Joe might talk down his interest in the bike path to avoid an obligation to ante up. Non-excludability lets Joe play the role of the free rider. Moreover, he is choosing to distort his preferences for the bike path network to avoid contributing, creating a serious problem for the economist who infers Joe’s value of the network of bike paths from his preferences (Frey et al., 2004; Buchanan, James M. Musgrave, 1999). The ‘free rider’ problem is one argument for government provision of certain goods by way of taxes (i.e. it seems a little unfair for Sally to have to foot the bill for a bike path network that Joe will also enjoy) but still the difficulty of ascertaining each
Sentiment analysis of tweets

individual’s demand function complicates determination of the optimal tax amount (Buchanan, Musgrave, 1999).

Over the years economists have invented a number of valuation techniques to circumnavigate the market failures induced by the non-rivalry and non-excludable nature of public goods. Nevertheless it is a difficult challenge and the current methods are not without their shortcomings. In light of these, another valuation method has emerged referred to as the life satisfaction approach (B. Frey et al., 2004). In recent years, economists have begun appropriating more attention towards happiness and the economics of happiness, thereby ensuing exploration into the implications of happiness data for environmental valuation models (Frey & Stutzer, 2002)(Welsch, 2009). What follows is a review of the weaknesses of the existing valuation techniques and thoughtful adaption of the life satisfaction approach to the advent of twitter data. Valuation approaches fall under two general categories, stated preference and revealed preference. The hedonic price and travel cost approaches fall under the category of revealed preference while contingent valuation and willingness to pay models are categorized as stated preference.

**Stated Preference Methods**

Contingent valuation models rely on surveys. They ask respondents to value a specific public good such as a bike lane (Loomis et al. 2000). Consequently, the validity of the valuation is subject to the cognitive capacity of the respondent to fully consider his or her preferences regarding the particular public good. The task is especially cognitively demanding given the respondent’s probable lack of familiarity with the public good and doubtful ability to quickly and accurately, recall and assess relevant budget constraints and alternative substitutes (Frey et al., 2004)(Welsch, 2009). Evidence of incomplete and unstable preferences in the case of contingent valuation has been observed and referred to as the “embedding effect” (Kahneman & Knetsch, 1992). Kahneman and Knetsch demonstrate that variation in the nature of the public good is not reflected through changes in the respondent’s willingness to pay (WTP). Moreover, respondent’s WTP varied systematically with the sequence of questions in the contingent valuation surveys, a seemingly irrelevant alteration. To some degree, these shortcomings can be mitigated by providing a credible payment system or presenting more information (Portney, 1994). Finally, given the hypothetical nature of contingent valuation surveys and the respondents’ uncertainty about how the responses might affect them in the future; they are
especially vulnerable to strategic replies. That is, “individuals have no incentive to disclose their true demand for non-excludable goods; it is advantageous to understate demand when it positively affects contribution requirements and to overstate demand otherwise” (i.e. Joe wanted Sally to foot the bill) (Frey et al., 2004 p. 3).

**REVEALED PREFERENCE METHODS**

The *hedonic market approach* conjectures that since individuals benefit from local public goods they favor living in areas with greater access to such public goods, and hence, bid down wages and bid up housing rents in those areas (Frey et al., 2004). The inseparability of many public goods from location makes it possible to consider markets that deal in location such as housing markets, surrogate markets for the public good itself. However the conditions under which traditional market forces prevail must still be met, and here lies the problem. The conditions are: that households must possess a very high level of information, there is near perfect competition in the market, prices adjust quickly, moving and transaction costs are minimal, and market restrictions are absent. Clearly, these conditions are extremely strict. (Frey et al., 2004 citing Freeman 2003)

The *travel cost or tourism approach* rests on the premise that people spend time and money traveling to different places for the purpose of consuming the public goods provided at those places. Public goods being valued under this scenario tend to be cultural institutions and recreational sites. One example is a study which evaluated recreational trout fishing damages in Montana’s Clark Fork River basin by assessing the travel costs accrued by trout fishers having to travel farther following contamination of nearby rivers with heavy metal mining waste (Morey et al. 2002). Individuals’ expenditures are likened to their WTP. Unfortunately the travel cost method is limited by the fact that consumption of the publicly provided good must entail costs in some way (Frey et al., 2004).

In sum, both stated preference and revealed preference methods are incomplete solutions to the challenge of public good and environmental externality valuation. Each rely on an individual’s conscious awareness of the effects of the public good or bad in question on their life, a condition that is often absent (Welsch, 2009). The life satisfaction approach is not subject to this constraint.

**LIFE SATISFACTION APPROACH**
The Life Satisfaction Approach evolved to a great extent from Daniel Kahneman and the late Amos Tversky’s research distinguishing two types of utility: *decision utility* and *experienced utility* (Kahneman et al., 1997). Their framework posits conventional economic utility theory as decision utility, stating “utility of outcomes and attributes refers to their weight in decisions: utility is inferred from observed choices and is in turn used to explain these choices” (Kahneman et al., 1997 p. 375). In contrast *experienced utility* is described as the hedonic experience associated with an outcome. This interpretation of utility actually came before the current notion of utility, belonging to the work of Jeremy Betham in his 1789 book *A Introduction to the Principles of Morals and Legislation* in which he considered pain and pleasure as the “sovereign masters” of whose authority it is “to point out what we ought to do, as well as determine what we shall do” (Betham 1789 p. 4). Why then, until recently did decision utility get all the attention? Kahneman, Wakker and Sarin argue the Bentham-ite grasp of utility was neglected for two reasons: 1. people’s hedonic experience is unobservable, immeasurable, and subjective and 2. all information is provided by choices. Since economic agents act on consistent preferences and rationally maximize utility.

With regard to the second objection, the field of behavioral economics has demonstrated that the preferences of individuals are not always well-defined, consistent or maximized (Warren, McGraw, & Van Boven, 2011). Methods for observing and measuring the subjective hedonic experience of an individual are still in development, but no longer considered impossible. The current study posits Twitter is a potential option. The life satisfaction approach, on the other hand, came about from the idea that subjective well-being survey responses are a way to measure experienced utility. If experienced utility is directly measurable, then the only prerequisite for determining the marginal rate of substitution between two goods is a regression model. Thus it is possible to ascertain the marginal rate of substitution between income and quantity of a public good. All the life satisfaction approach is a regression model where the metric for experienced utility is predicted by: A. a variable for income, B. a variable for the public good, and C. a vector of control variables (e.g. demographic indicators and socio-economic variables; Frey et al., 2004). In the past, most life satisfaction approach studies have used subjective well-being survey responses as the experienced utility metric, hence the name ‘life satisfaction’. The current study uses sentiment analysis of Tweets as the metric for experienced utility.
METHODODOLOGY

TWEET COLLECTION

Tweet collection for this project was conducted between January 6, 2013 and February 2, 2013. Tweets were accessed through Twitter’s application programming interface or API, which is free to the public. Twitter’s intended purpose for the release of this data is to make it easier for application developers (e.g. – makers of iPhone apps) to incorporate Twitter into their applications. However, a significant bonus is its relevance for academic and commercial research when combined with natural language processing.

Twitter has two options for tweet collection, search API and streaming API. Streaming API, likely represents the most thorough and reliable option, however it requires resources beyond the means of this project (i.e. - 24/7 servers). Streaming API accesses the Twitter server directly, and collects tweets returned by the server while the connection is maintained. This means that to collect tweets over an extended period of time, say a month, the data collector must have access to a server, which can maintain such a connection to continually collect data. Since the entirety of the current project was conducted on a personal laptop; using the streaming API was not an option. Using the non-streaming (Search) API has a few notable weaknesses, namely the Tweet author’s followers and friend counts, and the amount of tweets produced by the author since joining Twitter cannot be accessed. For the purpose of the current project, that information is not imperative. However, the current study would have benefitted from continuous collection and the even larger number of collected tweets that would have been afforded by the Streaming API.

Nevertheless, the search API is still a useful and plentiful source of data. The format of search API is an HTTP GET request, which retrieves REST URL, which calls on the search API. In practice this is actually quite simple, a query looks a lot like most of the URLs one enters into a web browser, except within it a number of user-configurable parameters can be specified. One could in fact, enter a custom query into their browser and be returned results in very little time (< 5 minutes). However, results are returned in json format, one that is not very “human-readable” (nice to look) or easy to use for amateur computer scientists. Furthermore, if one is looking to administer multiple queries, re-entering them into the web browser is
Sentiment analysis of tweets

inefficient. This process was automated in the current study with use of the programming language, Python. Python is free and open source with a large selection of user-contributed packages, making it an excellent tool for the project at hand.

As alluded to earlier, search API permits and requires the specification of some parameters. The parameter options include: query term(s), geocode, language, number of pages, result type and number of tweets returned per page (rpp). The query must include a query term, geocode or both. To ensure random selection of tweets from each city, only geocode was specified in this project (GPS coordinates). Alternatively, many of the studies mentioned earlier, which, sought to predict elections were only concerned with tweets related to the elections and hence would include candidate and party names as query terms (e.g.- “republican” or “Barack Obama”). The geocode specification includes a set of longitude and latitude coordinates, and radius distance. If the geocode is specified then the search will only return geo-tagged tweets, which fall within the designated radius. A tweet is geo-tagged if it is sent from a device with “location services” turned on.

This project collected tweets from 368 metropolitan statistical areas as defined by the Office of Management and Budget. The GPS coordinates associated with each those cities were scraped from Google API using the geopy package of Python. Finally, in order to capture all tweets coming from each of the cities, a 20-mile radius from the center of the city was specified.

The other parameters will now be discussed. Only tweets written in English were collected. Result-type has 3 options, mixed, recent and popular. Recent returns only the most recent tweets posted, popular returns only the most popular tweets posted and mixed returns a mixture of recent and popular (twitter website citation). Here, recent is specified to minimize any bias. To maximize to possible number of tweets returned for each search, number of pages is set at 15 and tweets returned per page is set at 100. Each of those parameters is the maximum allowed; meaning the most tweets a single query can collect is 1500.

When search API queries retrieve tweets it accesses tweets stored on Twitter’s servers. Twitter’s servers only save tweets for a maximum of 5-7 days and sometimes less when the current load is high (e.g. – New Year’s Eve). This stipulation has particular implications for research planning and the scope of the current project. Tweets cannot be collected from the distant past and the desired timespan of the tweets dictates the
length of the project and vice versa. Notably, while the largest possible dataset is the most desirable, the current study was limited to collecting tweets when time allowed and only through HTTP GET requests. That being said, the dataset for the current study is enormous (discussed below).

As was noted earlier, to facilitate the repetition of conducting 368 queries (one for each metropolitan area), a Python script was written. The script was run every couple of days starting in January 2013. As of the most recent collection, the total number of tweets collected in the US was 1,864,063. Text, time, date, user id, and geocode were recorded. Naturally, given the specified GPS coordinates in the geocode search parameter, the city, state and country were also available.

**Sentiment Analysis of Tweets**

The most basic idea behind automating sentiment analysis of text is to identify features of the text that are indicative of its polarity. Features of text can take a large range of forms; some examples are: words, two words that frequently appear next to each other (bigram), and the length of the text. A study concerned with automatically assessing the quality (not polarity) of Wikipedia articles, involved some 50 features in their classifier. These features included length measures, such as, number of paragraphs and number or words, part of speech measures, which are essentially counts of different parts of speech in the article, and web-specific characteristics, for instance, the use of images and external links (Rassbach, Pinock, & Mingus, 2007). Analyzing shorter texts like tweets presents a distinct challenge because their length limits how many features can be employed.

One approach is to focus on the words in the text and compare them to lists of words, which have been pre-selected as positive or negative. Each tweet is cross-referenced with the list of positive words and the list of negative words and a tally is kept for each word match. At the end there is a count of how many positive words and/or negative words were present in each tweet. If only positive words or more positive than negative words were found in a tweet then the logical conclusion is that the tweet is positive. Further refinement might be some measure of how positive or how negative the tweet is. With that in mind, one might suppose that proportional to the length of the tweet, the more positive words present in the tweet indicates the degree of positivity of the tweet. Another consideration might be that some words express greater positivity or greater
Sentiment analysis of tweets

negativity than others. To address this challenge one could add rankings to the words in the list of positive and negative words. The lexicon utilized by (Mitchell et al., 2013) scored each word on a scale of 1 (sad) to 9 (happy). For each word in a tweet that corresponds to a word in the lexicon the happiness score is recorded and the total happiness score is calculated.

**Sentiment Lexicons**

To implement the methodology above, one must start with a pre-defined lexicon. There a quite a few affective word lists that have already been created that vary in their accessibility. The Linguistic Inquiry and Word Count (LIWC) was one of the first sentiment dictionaries to be developed and is now available commercially in conjunction with a software program to facilitate the text analysis (Pennebaker et al. 2007). Three versions have been released since its original conception. The Harvard General Inquirer is very similar to the LIWC and is available for academic purposes by request. Originally conceived in 1966, like the LIWC a good amount of time has been spent on its development; although it has not been updated in some time (Stone et al. 1966). One of the more popular word lists used for sentiment analysis of Tweets is, Affective Norms for English Words (ANEW), originally released in 1999 prior to the advent of micro blogs (Bradley & Lang, 1999) The original list of words and their emotional valance was published in the initial publication, however, the most up to date version is only available by request for academic purposes and can take up 30 days to acquire (http://csea.phhp.ufl.edu/media/anewmessage.html).

Alternatively, AFINN is a wordlist, which has been shown to perform comparatively to, if not slightly better than, ANEW, and is distributed freely via the internet (Nielsen, 2010). Having compiled the word list specifically for the analysis of tweet chatter surrounding the United Nation Climate Conference (COP15) and future tweet sentiment analysis, the author aspired to include words relevant to the vernacular of micro blogs, such as, obscene language and internet slang. The words are scored for polarity between -5 (very negative) to +5 (very positive), the majority of, positive words labeled +2 and negative words labeled -2. The current study intends to use the AFINN affective word list to facilitate one of its tweet sentiment analyses.

Also publicly available is the MPQA Subjectivity Lexicon part of the OpinionFinder software, which was produced by Theresa Wilson, Janyce Wiebe, and Paul Hoffmann at the University of Pittsburgh.
Sentiment analysis of tweets

(Wilson, Wiebe, & Hoffmann, 2005). There are 3095 words in the negative list and 2230 words in the positive list.

Finally, continuing with the lexicon methodology, an increasing number of studies take advantage of the Amazon Mechanical Turk service (Mitchell et al., 2013; Dodds et al., 2011). Amazon Mechanical Turk is an Internet based crowdsourcing service, which coordinates the completion of tasks, which require basic human intelligence for those in need of such, for example the scoring of words as positive or negative. Users of Amazon Mechanical Turk can work on tasks at their leisure and receive small payment for their contribution. Meanwhile, companies and researchers can submit tasks such as, scoring the affective polarity of 10,000 words or the polarity of the tweets themselves, and pay a small fee for the task’s completion. This is a popular way in sentiment classification to test a classifier; have humans manually classify a portion of the data and test the accuracy of the classifier on it. In the current study, we use a lexicon that has been validated by humans via Mechanical Turk as well as comparison to other lexicons.

DATA SET

The dependent measure for this project is affect levels of cities estimated through analysis of collected tweets. Independent variables for this project are data provided by various organizations and government agencies. Those organizations and agencies are: the American Community Survey, Bureau of Labor Statistics, American Transportation Association, Federal Bureau of Investigation, and the League of American Cyclists. The bulk of the independent data was extracted from the US Census Bureau, 2011 American Community Survey (ACS) 1-year estimates for all US metropolitan statistical areas. From the ACS, the following subsets of data were selected: educational attainment (S1501), commuting characteristics (S0801), selected economic characteristics (DP03), and Gini index of income inequality (B19083). The measures for unemployment for each city were sourced from the Unemployment Rates for Metropolitan Areas (monthly rankings not seasonally adjusted, Jan. 2013), a data set made available by the United States Department of Labor, Bureau of Labor Statistics. The measures of public transportation usage and availability were extracted from the 2012 Public Transportation Fact book (Appendix B: Operating Statistics), a data set made available by the American
Sentiment analysis of tweets

Public Transportation Association. The report was compiled from data reported by transit agencies to the Federal Transit Administration’s 2010 National Transit Database. The measure of crime presence within cities was extracted from the Federal Bureau of Investigation’s (FBI) Uniform Crime Reporting Program, specifically the Preliminary Semiannual Uniform Crime Report (January - June 2012). The measure of bike friendliness/accessibility of each city was extracted from the 2009 Lane and Path Mileage data compiled by the League of American Bicyclists.

To construct a full data set to run regression models (and other analyses) on, the program R was used to match rows of data from the same city and merge together all data points. Because each data source did not use the same geographic unit definition, the combined data set did not include every US city. Tweets were collected based on the Metropolitan Statistical Areas (MSA) as defined by the Office of Management and Budget (OMB). The boundaries of MSA’s often groups neighbor cities and towns into a single MSA (e.g. –Washington-Arlington-Bethesda). In cases such as these the most populous or leading city was taken to represent the area. This was necessary because twitter search queries require a single GPS coordinate input. American Community Survey data was also organized by MSA’s. Unemployment data was roughly organized around MSA’s although there was some divergence. Public transportation data was organized around urbanized areas (UZA). UZAs are census designated urban areas with populations of at least 50,000. This definition is in contrast to MSAs, which are an urban area of at least 50,000 as well as one or more counties containing the urban area and adjacent counties, which are highly socially and economically integrated with the core urban area. Hence the MSA represents a larger area than the UZA and in some cases groups multiple UZAs together. Crime data was reported only for those cities with population of 100,000 or more. Finally the bike lane and path mileage was only available for the 90 largest US cities, which did not all line up with MSAs.

Faced with the tradeoff between more cities or more variables of interest, the final analysis data set converged on variables that were available for at least 360 MSAs. A separate data set was constructed for follow-up/future analyses that contained all variables, but with only 60 cities represented (see Future Directions). For all reported analyses, the number of cities was n=361.
Sentiment analysis of tweets

**Tweet Analysis Overview**

A simple text analysis was run which cross-referenced the AFINN-111 lexicon against the tweets (parsed into single words). The structure of the AFINN lexicon is as follows: words are rated for valence between -5 (negative) and +5 (positive). The author of the AFINN lexicon compiled the list from various relevant affective word lists, obscene word lists, slang from Urban Dictionary, and clustering analysis of tweets to extract relevant words and then manually rated by Finn Årup Nielsen in 2009-2011. For a thorough explanation of the construction of the word list see (Nielsen, 2010). Validation of the lexicon was conducted by comparing its ratings to those obtained using Amazon Mechanical Turk and other established sentiment classifiers and lexicons. The author decided to use this lexicon in the current study because the valence ratings were novel compared to other lexicons. Additionally, the language of the lexicon seemed most relevant to language used in tweets. Many lexicons such as the MPQA lexicon were developed to assess much longer and more formal texts, which tweets are not. Finally, the author modified the AFINN-111 lexicon to include emoticons. [ :) ( : ) ( : -D : p ; ) B ) : ] were added and given a valence rating of 5 (positive). [ : ( : -{ ; ) ; : ] were added and given a valence rating of -5 (negative).

To extract the final measure of sentiment from the tweets, each tweet was parsed into single words. For every word in the tweet that matched a word in the AFINN-111, the corresponding valence rating of the word was recorded. These values were summed together for each tweet and divided by the total number of matched in the tweet. For a tweet that contained the word bad (-3) and emoticon : (5), the total affect score for the tweet would be 2. This method was similar to the method used in another study of tweet affect (Mitchell et al., 2013). A key difference between the current study and that study is the nature of the rating in the current study being positive and negative rather than sad and happy. Additionally, the current study used a range that centered around 0 rather than the range 1 to 9.

---

3 Emoticons are thought to be highly indicative of the sentiment of tweets. (Kouloumpis, Wilson, & Moore, 2011) use a corpus of tweets, which contain either a positive emoticon or negative emoticon to assist in training an AdaBoost.MH (Schapire and Singer 2000) supervised learning classifier.
To compute a single tweet affect score for each city, the tweet affect scores were averaged. Descriptive statistics for the tweet affect score across all cities are as follows: mean= 0.265 (n = 361) and the median= 0.262 (n = 361). At first blush, the mean tweet affect score might appear very low, however this is not unexpected. A tweet affect score of zero could mean one of two things, either all-positive word scores canceled out the negative words scores or there were no AFINN-111 word matches. Both of these possibilities are fairly indicative of neutrality of the tweet. However average frequency or word matches per tweet aggregated by city is 0.867 (n = 375) suggesting that a portion of tweets had no AFINN-111 word matches. When analyzing such a large data set (approximately 1.8 million tweets total) not every tweet is expected to have a word match with the AFINN-111 lexicon or for every tweet to express sentiment.

\[ \text{mean} = 0.265 \quad \text{(n} = 361) \]
\[ \text{median} = 0.262 \quad \text{(n} = 361) \]

4 Not overall average for cardinality reasons.
RESULTS

DATA INTEGRITY

When using a novel approach such as sentiment analysis of tweets, it is important to verify your measures as actually measuring the construct of interest. Additionally, for any follow-up analyses using these measures, it is important to verify variability between cities. That is, if cities do not differ in affect measures, or explanatory variables, follow-up analysis will be flawed or not possible.

Figure 1 shows the distribution of tweet affect score. There is sufficient variation in tweet affect score across the cities. Furthermore, the distribution of tweet affect score is approximately normal after removal of outliers\textsuperscript{5}. Also of note, the distribution sits almost entirely to the right of zero indicating that tweet affect score means of cities were mostly positive. Tweet affect scores ranged from -0.110 to 0.560. This is especially interesting because the AFINN-111 lexicon is slightly skewed toward negatively rated words with 1598 negative words (about 65\%) compared to 878 positive words. In spite of this skew the distribution of mean tweet affect scores of cities is primarily within the positive range.

\textsuperscript{5}Bremerton, WA and Barnstable Town, MA were considered outliers for having mean tweet affect scores greater than six standard deviations away from the mean.
Figure 1. Variability across cities in mean tweet affect score.

Another concern for the integrity of the data set is whether the same number of tweets was collected from all of the observed cities. For each Twitter search query a maximum of 1500 of the most recent tweets could be collected, but there was no guarantee of this happening for every city in every search query. Figure 2 shows the distribution of the number of tweets collected for each city. Unfortunately there is considerable variation. In order to compensate for this variation, it will be controlled for in the regression models.
The final data integrity step was to investigate whether or not tweet affect aggregated by city can serve as a proxy measure of experienced utility. To check whether mean tweet affect score is a valid measure of experienced utility, the experimenter performed Pearson's product moment correlations between tweet affect and variables for which the experimenter had a priori predictions. The variables for the correlations were chosen from intuition and literature concerning the measurement of experienced utility. Table 2 displays the results of the correlations and whether they followed the predicted direction. The predicted direction was confirmed for all the variables except those pertaining to income.

Table 2.
Sentiment analysis of tweets

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Variable</th>
<th>Predicted Direction</th>
<th>Correlation</th>
<th>Check</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.</td>
<td>% pop. 25 &amp; up with Bachelor’s Degree</td>
<td>positive</td>
<td>0.219**</td>
<td>✓</td>
</tr>
<tr>
<td>b.</td>
<td>% pop. 25 &amp; up with Graduate or Professional Degree</td>
<td>positive</td>
<td>0.154**</td>
<td>✓</td>
</tr>
<tr>
<td>c.</td>
<td>% pop. 25 &amp; up High School Drop outs</td>
<td>negative</td>
<td>-0.216***</td>
<td>✓</td>
</tr>
<tr>
<td>d.</td>
<td>% Families &amp; People Living Below Poverty Line</td>
<td>negative</td>
<td>-0.097</td>
<td>✗</td>
</tr>
<tr>
<td>e.</td>
<td>Mean Household Income</td>
<td>positive</td>
<td>0.030</td>
<td>✗</td>
</tr>
<tr>
<td>f.</td>
<td>Unemployment Rate</td>
<td>negative</td>
<td>-0.179***</td>
<td>✓</td>
</tr>
</tbody>
</table>

(p<0.001) ***, (p<0.01)***, (p<0.05)*

**Figure 3** presents histograms of the key independent variable incorporated into the higher-level analyses. For the most part, the distributions appear relatively normal. The distribution of the percentage of the population that commutes to work by bicycle is heavily skewed to the right. The distribution of the percentage of population self-employed also diverges slightly from the normal distribution. The shape of the histogram possibly suggests a Fisher Distribution.
Figure 3. Histograms of Key Independent Variables

- Population 25 years or Older with Bachelor’s Degree
- Percentage of Population that Commutes to Work by Bicycle
- Median Earnings of Population 25 years and Older
- Mean Travel Time to Work
- Percentage of Population Without Health Insurance Coverage
- Unemployment Rate
Sentiment analysis of tweets

**REGRESSION ANALYSES**

The author ran several regression models to address a series of hypotheses.

**EXPLORATORY HYPOTHESES BASED ON A PRIORI PREDICTIONS**

1. Age of the population of the cities will influence tweet affect score. Specifically, we expect cities with larger percentage of young citizens to have higher tweet affect scores.

2. Cities whose citizens have shorter commute times to work will have higher tweet affect scores. The experimenter postulates that shorter commute times might be indicative of cities with a greater sense of community resulting from proximity.

3. Cities with a higher percentage of bike commuters will have higher tweet affect scores.

4. Cities where a greater percentage of self-employed workers will have higher tweet affect scores.

5. Cities with a greater percentage of population living without health insurance coverage will have lower tweet affect scores.

6. Cities with proportionally more Bachelor’s Degrees will have higher tweet affect scores.

7. Cities with higher median income will have higher tweet affect scores.

8. Cities with lower unemployment rates will have higher tweet affect scores.
Sentiment analysis of tweets

<table>
<thead>
<tr>
<th>Table 3. Regression Model Parameters and Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OLS Regression of Mean Tweet Affect Score per City</strong></td>
</tr>
<tr>
<td><strong>Model A</strong></td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>Population Age 18 - 24 years</td>
</tr>
<tr>
<td>Population Age 25 - 34 years</td>
</tr>
<tr>
<td>Population Age 35 - 44 years</td>
</tr>
<tr>
<td>Population Age 65 and up</td>
</tr>
<tr>
<td>Travel time less than 10 min</td>
</tr>
<tr>
<td>Travel time more than 60 min</td>
</tr>
<tr>
<td>% Self Employed</td>
</tr>
<tr>
<td>% Commute by Bicycle</td>
</tr>
<tr>
<td>% Population with B.A. Degree</td>
</tr>
<tr>
<td>% Without Health Insurance</td>
</tr>
<tr>
<td>Unemployment Rate</td>
</tr>
</tbody>
</table>

All of the regression models controlled for the total number of tweets collected from each city. This control variable was used because of variability among the total number of tweets collected from each city (i.e. – some cities had relatively few (1000-3000) tweets collected). The total number of tweets from each city did not significantly predict mean tweet affect score in any of the 9 models.

In the first model, Model A, tweet affect of a city is predicted by the age distribution of the population; broken up into brackets. Of the five age brackets in the model, ranging from 18–24 to 65+, only two are significant. The beta coefficient for the percentage of the population aged 25 to 34 was positive and significant, while the beta coefficient for the percentage of the population aged 45 to 65 was negative and significant (see Table 3A for coefficients and t-scores). Both these results provide evidence for the a priori prediction that on average, younger populations will be associated with higher tweet affect.
The second model, **Model B**, maintains the age variables as controls and adds travel time to work variables (from S0801 Commuter Characteristics data set) as predictors. Commute time data was available in brackets. Only two variables, the tails (travel time less than 10 min and travel more than 60 min) were chosen for the models. Of the commute time variables, only the left tail (travel time less than 10 min) was significant (see **Table 3B** for coefficients and t-scores). The nature of the significant relationship was such that greater proportion of the population of a city with a commute to work of less than 10 minutes was associated with an increase in the mean tweet affect score of a city. Additionally, the significance of the coefficient for the percentage of the population aged 25 – 34 was maintained (as in Model A).

The third model, **Model C**, maintains the age and commute time variables as controls and adds a class of worker variable (from DP03 Selected Economic Characteristics data set). The class of worker variable is the percentage of the population that reported being self-employed in own not incorporated business workers. It does not require that the person be currently employed or in the labor force, only that they reported having a job in the last 5 years. It encompasses people, who worked for profit or fees in their own unincorporated business, profession or trade, or who operated a farm, as described by the US Census Bureau. The coefficient for the percentage of population self-employed was positive and significant (see **Table 2C** for coefficients and t-scores). The nature of the significant relationship was such that increases in the proportion of the population that reported being self-employed in their most recent occupation was associated with an increase in the mean tweet affect score of a city. The significance for the beta coefficients for the percentage of the population aged 25 -34 (as in Model A) and percentage of commutes of travel time less than 10 minutes (as in Model B) continued as well.

The fourth model, **Model D**, preserves the age, commute time, and class of worker variables as controls and adds another variable from the commuter data set (S0801). The new variable is the percentage of the population that commutes to work by bicycle. The beta coefficient for the percentage of the population that bike to work was positive and significant (see **Table 3D** for coefficients and t-scores). The nature of the significant relationship is such that in cities where a greater proportion of the
Sentiment analysis of tweets

population arrive at work by bicycle is associated increases in mean tweet affect per city. Additionally, the significance of beta coefficient for the total population aged 25 -34 (as in Model A) carried through but the beta coefficient for travel time to work less than 10 minutes (as in Model B) did not maintain significance in Model D. Additionally, the significance of beta coefficients for the percentage of population self-employed dropped out.

The fifth model, Model E, maintains the age, commute time, class of worker and bicycle commuter variables and introduces estimated population 25 years and older with a Bachelor’s Degree (from S1501 Education Attainment data set). The education attainment data set groups the estimated population with a Bachelor’s Degree by age brackets. 25 years and older chosen as most representative of the city as a whole, given that most Bachelor's Degrees are earned before the age of 25, and hence if the people in the sample intended to obtain Bachelor's Degrees we would expect them to have done so already. The beta coefficient for estimated population with a Bachelor’s Degree was positive and significant (see Table 3E for coefficients and t-scores). The nature of the significant relationship is such that as the total estimated population with a Bachelor's Degree increases, the mean tweet affect score increases as well. Additionally, the beta coefficients for, the percentage of the population aged 25 – 34 (as in Model A), travel time to work less than 10 minutes (as in Model B), and the percentage of the population that commutes to work by bicycle (as in Model D) all maintained significance in Model E. In this model, the population aged 45-64 variable regains significance.

The sixth model, Model F, maintains the age, commute time, class of worker, bicycle commuter, and education variables. Model F introduces a variable that indexes health insurance coverage (from DP03 Selected Economic Characteristics data set). The new variable is the percentage of the population with no health insurance coverage. The beta coefficient for the percentage of the population with no health insurance coverage is positive and significant (see Table 3F for coefficients and t-scores). A positive association between percentage of population with no health insurance and mean tweet affect score was unexpected. Additionally, the beta coefficients for, the percentage of the population aged 25 – 34 (as in Model A), travel time to work less than 10 minutes (as in Model B), and the percentage of the
Sentiment analysis of tweets

population that commutes to work by bicycle (as in Model D), and population 25 years and older with a Bachelor’s Degree (as in Model E) all maintained significance in Model F.

The seventh model, Model G, preserves the age, commute time, class of worker bicycle commuter and education variables, and percentage of population with no health insurance and adds the unemployment rate (from US Bureau of Labor Statistics data set). The beta coefficient for the unemployment rate was negative and significant (see Table 3G for coefficients and t-scores). The nature of the relationship is such that as the unemployment rate decreases in cities, the mean tweet affect score increases. Additionally, the beta coefficients for travel time to work less than 10 minutes (as in Model B), the percentage of the population that commutes to work by bicycle (as in Model D), population 25 years and older with a Bachelor’s Degree (as in Model E), and percentage of population with no health insurance coverage (as in Model F) all maintained significance in Model G. In Model G, no significant beta coefficients were found among the age variables.

The eighth model, Model H, and ninth model, Model I, are nearly the same. The final variable introduced is the median earnings for the population 25 years and older (from DP03 Selected Economic Characteristics data set). Both models contain all the variables included in Models A through G, except Model I leaves out the percentage of population self-employed variable based on its lack of significance following its initial introduction and possible issues of multicollinearity stemming from the correlation between it and the health insurance variable. The beta coefficient for median earnings and the population 25 years and older were negative and significant (see Table 3H and Table 3I for coefficients and t-scores). A negative relationship between median income age 25 and up and tweet affect score was unexpected. Significance for the beta coefficients for travel time to work less than 10 minutes (as in Model B), the percentage of the population that commutes to work by bicycle (as in Model D), population 25 years and older with a Bachelors Degree (as in Model E), and percentage of population with no health insurance coverage (as in Model F) all maintained significance in Models G and I. In both models, no significant beta coefficients were found among the age variables. Notable between the
Sentiment analysis of tweets

two models is **Model I**, which leaves out the percent of population self-employed variable, has a slightly larger adjusted r squared value than **Model H**.
DISCUSSION

The experimenter first considered whether tweet affect aggregated by city was a valid measure of the desired construct, *experienced utility*. Checks of whether tweet affect scores are measuring experienced utility (or the affective state of a city) can be conducted by correlating the measure with other variables for which we have a priori predictions. For instance, we generally consider there to be a positive relationship between higher education levels and quality-of-life/well-being. We would also expect a negative relationship between unemployment and tweet affect scores. Checks such as these have also been used as evidence of the validity of subjective well-being surveys (Kahneman & Krueger, 2006). One study confirmed that such a check is possible when they found a relationship between unemployment and subjective well-being surveys (Tella et al., 2007). In the same vein, the well-known “Misery Index” is simply the sum of the unemployment rate and inflation rate, as both are considered to deteriorate economic well-being. *Table 1* shows the relationship between tweet affect and variables the experimenter chose for their association with higher quality-of-life to validate tweet affect. In this check, the experimenter observes significant and directionally correct relationships among education attainment measures and unemployment rates. So, we conclude that tweet affect seems to be a reasonable proxy measure for experienced utility. This conclusion is one of the most significant contributions of the current study due to the relatively low computational and financial overhead associated with collecting and analyzing tweets. Whereas measures of experienced utility are typically costly and time consuming to collect, tweet collection, analysis, and aggregation are cheap and able to be automated.

After determining that tweet affect scores are a reasonable proxy for experienced utility, several higher-level hypotheses were formed and tested with multiple, multiple regressions. The first of those hypotheses was that cities with on average younger populations would score higher in tweet affect. This was an intuitive prediction based on the commonplace assumption that younger cities are seen as more desirable. The opposite direction of this prediction is that cities with a higher proportion of older citizens would have lower tweet affect scores. In addition to this linear type relationship between age and affect, non-linear relationships may be predicted. In fact (Stone et al., 2010) studied responses to life
Sentiment analysis of tweets

evaluation questions across a sample of people aged 18 to 85 years. They found a “U bend” relationship. That is, on average, people are happy in early years of life. As they approach middle age this happiness declines, but after early fifties it begins to rise again, hence the “U bend”. While this study investigated the relationship between affect and age in individuals, the current study addresses the question in a cross-sectional analysis of cities. The results of Model A (see Table X) suggest a “U bend”, though shifted slightly to the right. Of the age brackets in Model A only two were significant, *percentage population aged 25 to 34* and *percentage population aged 45 to 64*. The coefficient on total estimated population aged 25 -34 is positive and the coefficient on age 45 to 64 is negative. The directionality of these relationships is in agreement with our hypothesis and the study that found a “U Bend”. 25 -34 is still in the happy range at the beginning of the U and 45 - 64 hovers around the trough of the U. The coefficient on 65 and up is in the predicted direction (positive) but it is not significant, while the coefficient on 18 - 24 is negative but also extremely insignificant. So, there is evidence that the prediction made in the first hypothesis was reasonable. Furthermore, the results of Model A do not differ greatly from those in (Stone et al., 2010) which predicted global life satisfaction survey evaluation; providing further validation of tweet affect score being a proxy for experienced utility.

In Models B through I, we introduced variables selected from the combined data set (1012 variables observed for 359 cities) that we hypothesized would be related to tweet affect scores:

In Model B we hypothesized that cities where commute times were shorter would be associated with higher scores in tweet affect, and conversely cities with longer commute times would score lower in tweet affect. Shorter commute times might free up more time for the city's population to devote to leisure such as spending time with family. We also thought that shorter commute times might indicate that in addition to the workplace, other socio-economic and entertainment qualities might also be grouped near the residence, potentially contributing a greater sense of community. Hence an indirect hypothesis is that cities more oriented towards the community would score higher in tweet affect. A variable to indicate extremely short commute times and one to represent the opposite end of the spectrum, very long commute times, were introduced in model B. The beta coefficient for the commute
Sentiment analysis of tweets

time less than 10 minutes was significant and in the direction we predicted, positive. On the other hand, the variable indicating on average longer commute times (60 min plus) was not significant. So, there is evidence of our prediction that very short commute times are associated with more positive tweet affect. But the effect of very long commute times is still unclear. It may be that another U shaped relationship rather than strictly linear is present in the population. It might be that cities with larger proportions of the population living over an hour away from their place of work choose to do so based on the presence of interesting land features outside the city or a desire to avoid the hustle and bustle of the city, rather than living far away from their work for lack of a better alternative.

In **Model C**, a measure indicating the percentage of the population that was self-employed in the city was included. We hypothesized a positive relationship between the proportion of the population that is self-employed and tweet affect score. This prediction was driven by acknowledgement of some the freedoms associated with being your own employer such as flexibility of schedule and autonomy in work goals. The hype surrounding the startup culture in Boulder, and across the country, also motivated the question of whether a greater presence of self-employed workers in a city positively affects mean tweet affect score. Boulder's incubator TechStars was ranked second to the top in Tech Cocktail’s “2012 Top Fifteen USA Startup Accelerators” ([http://tech.co/top-startup-accelerators-ranked-2012-08](http://tech.co/top-startup-accelerators-ranked-2012-08)). The beta coefficient for the percentage of the population self-employed was positive and significant in Model C, providing evidence in favor of our prediction. However, with the introduction of more variables in models D through G, this predictor drops out, indicating this finding is either not robust or suffers multicollinearity issues with one or more of the following variables.

In **Model D**, we hypothesize cities with more bicycle commuters will on average be associated with more positive tweet affect. As a bicycle commuter, the author has a personal investment in this prediction. However, in addition, there is substantial portion of research highlighting the health benefits of cycling ([Buehler & Pucher, 2011](http://tech.co/top-startup-accelerators-ranked-2012-08)). English writer, H. G. Wells, famously said, “Every time I see an adult on a bicycle, I no longer despair for the future of the human race.” The beta coefficient for the percentage
Sentiment analysis of tweets

of the population that commutes to work by bike was positive and highly significant, confirming our prediction. Furthermore the size of the effect is large relative to the data (see Table D).

In **Model E**, we continue with our hypothesis from the checks portions, that higher educational attainment is associated with more positive affect in tweets. Higher education and well-being have long been considered to be positively associated (Witter et al., 2013). For the individual, higher education increases access to higher paying and stable jobs and provides time for self-development and discovery. At the city level we might expect higher education would facilitate cooperation and compromise in cities and so, expect cities with a greater proportion of the population holding Bachelor’s degree to score higher in tweet affect. In the **Model E** we test this hypothesis by including a variable for the percentage of the population 25 years and older with a Bachelor’s degree, and observe a positive and significant beta coefficient. Once again we validate measuring experienced utility with tweet affect by linking higher education to positive affect in tweets, even after controlling for other variables such as age.

In **Model F**, we ask what is the effect of a greater percentage of the population living in a city with no health insurance. We predict a negative relationship between the proportion the population with no health insurance and tweet affect score as health insurance may provide a degree of internal comfort to the individual. This question is partially motivated by the political debate regarding the provision of health insurance happening in the USA currently. Surprisingly, the beta coefficient for the percentage of the population with no health insurance coverage is significant but negative, contrary to our prediction. We consider that this may be an effect that illustrates the differences between an analysis of individuals versus a cross-sectional analysis of data aggregated by city. While lack of health insurance may be discomforting for the individual, when looking at cities we might not be accounting for qualities of the cities which make up for a larger proportion of the population living without health insurance coverage. Cities with a historically high population living without health insurance may have developed programs and systems to compensate. For example, some cities such as Cleveland, Ohio have excellent hospital systems (Cleveland Clinic) that may provide services/assistance to those without coverage. To explore the effect of health insurance on tweet affect, a future model might include some
Sentiment analysis of tweets

measure of the quality of social services and alternative health care options of the city. Additionally, given the recent passing of legislature to provide health care to all Americans, a future study might conduct a before and after analysis of the effect of health insurance coverage on tweet affect scores in cities.

In **Model G**, we hypothesize that higher unemployment rates will be negatively associated with mean tweet affects scores of cities. This prediction is fairly intuitive. Cities with a higher unemployment have more individuals who are likely stressed monetarily (e.g. – less opportunity to do paid recreational activities). Both these consequences are detrimental to the well-being of society. As predicted, the beta coefficient for unemployment rate is negative and significant. Also with the introduction of unemployment rates into the model, the adjusted r squared jumps from 0.177 to 0.213.

In **Models H and I**, we hypothesize that tweet affect scores will increase with income. For a long time economic growth and high GDP were seen as a top priority by individual and governments. This is not necessarily the opinion of this study. The Easterlin Paradox was the first to show that over an extended period of time, subjective well-being responses in countries did not increase despite substantial increases in income level during that period of time (Easterlin, 1974). In a more recent study, both life evaluations and measures of emotional well-being of individuals were shown to steadily rise with income up an annual income of $75,000 after which rise in income no longer increased happiness (Kahneman & Deaton, 2010). The current analysis is looking at the median earnings (population 25 years and older) of cities. The maximum median earnings in the sample was $51,746. We predicted that increased income would be associated more positivity in tweets. However in both **Model H and Model I**, the beta coefficient for the median earning of the population at least 25 years was negative and significant. This is a fairly surprising result but provides further evidence that the relationship between income and happiness is very tricky. Unfortunately, a negative coefficient on income renders the Life Satisfaction Approach impossible. The ratio of the bike commuting coefficient and the income coefficient would be negative. This would imply that bike commuting is a public cost rather than good in spite of the positive relationship between bike commuting and tweet affect.
Across all models (A-H) we observed consistency in parameter significance with only minor exceptions. For example across all models we observe consistency in size, direction and significance in the coefficient for percentage of the population the commutes to work by bicycle. An example of an exception is the variable *percentage of population self-employed*. This parameter is only significant in **Model C**. This may be because, after adding control variables, the previously significant predictor no longer explains a significant proportion of the variance in the regressand. Another explanation is that the variable may suffer from multicollinearity with the added variables. Recognizing the inconsistency in the variable, percent of the population self-employed, we do not include it in **Model I**. We find that **Model I**, differing from **Model H** only by the absence of the dropped variable, actually explains slightly more of the variation in mean tweet affect score (adjusted r-squared 0.213 to 0.214).
Caveats and Future Directions

Statistical Caveats

We acknowledge that our models may have suffered from multicollinearity. A confusion matrix may have been helpful to identify potential multicollinearity issues before designing the models. However with many variables (1100+), a confusion matrix would also have been extremely difficult to interpret and likely require clustering or other data reduction techniques. Further attention could be placed on coming up with variable transformations and interaction terms to mitigate the multicollinearity issues when the variables are important to the theoretical predictions.

Aside from avoiding multicollinearity in models, there is also a potential to over fit the regression model given the size of the data set. We had access to nearly 3 times as many variables as cities. Hence careful consideration of what variables were actually of interest was necessary in this analysis and will be in all future analyses.

Finally, we took for granted that the mean tweet affect score was normally distributed from looking at its histogram. However, further investigation into whether this was a correct assumption would be appropriate.

Data Set Combination

As was mentioned earlier, we actually collected data from a few agencies in addition to the data sets acquired from the US Census American Community Survey Data. However, differences in the geographic units in the data sets caused substantial reduction in the size of the data set. Although we would have liked to explore the relationship between say, crime rates and tweet affect score, we would have had to sacrifice 100+ cities. Instead we chose to stick with the variables available through the American Community Survey data to maximize the number of cities observed. However, collection of the other interesting variable has been completed and could be introduced into models with less data points (60) in future analyses.
Sentiment analysis of tweets

Additionally, tweets and some independent variables were collected for the UK. Time limitations and computation challenges caused this data to be withheld from the current analysis. However, the UK could be an interesting country to extend the analysis to, especially due to the recent adoption of subjective well-being surveys in the UK. A future analysis could test the correlation between subjective well-being survey responses and affect in tweets and also compare models predicting each.

One take away from the experience of combining data sets, is that in the future, to maximize the data set with the most important independent variables, the independent variable data sets should be chosen before beginning the tweet collection. Say, for example, we were the most interested in crime rates, then we may want to collect tweets to match the geographic units and boundaries of the FBI preliminary report and forego the ACS data set.

In that same vein, the current study was limited by its ability to link tweets to cities precisely. The framework of the GET search Twitter API allows the specification to collect only geo-tagged tweets tweeted from a geographic area defined in the search. In the Twitter API query framework a specified radius around GPS coordinates defines search the geographic area. To link tweets to cities the current study specified the GPS coordinates of the center of the desired city and collected all geo-tagged tweets tweeted within a 20-mile radius. However, the 20-mile radius was chosen somewhat arbitrarily to, on average, match the area of the city. Consequently, for larger cities it may not capture all the tweets and for smaller cities it may grab tweets from outside the city. In future studies the specified radius should change for each city to more precisely reflect the boundary of the city.

A final limitation of the data set is the lags on the variables are not uniform. The tweets cannot be collected from the past and the independent variable data sets had to come from the past. We don’t think this substantially harmed our model, especially since we still found significant predictors. We assume many of the variables we used remained fairly steady for the duration of the lag. However, to avoid this issue the project would have to be designed in such a way that tweets were collected and then saved until the independent data sets were released.
**Hardware and Time Limitations**

While the data set for this project was quite large, it could have been larger and more complete. The collection of tweets for this project was limited by hardware and time. Twitter API has two options for collecting tweets, the GET search and Streaming API. The first, GET search, samples data that is stored on the Twitter server. Twitter allows a maximum of 1500 tweets to be collected in a given query. Queries were run every couple of days, but Tweets accumulate faster than is possible to keep up with. Queries would have likely continued collecting new tweets had they been run hourly. Tweets have a maximum lifespan of 7 days on Twitter’s servers, before they are deleted and replaced with incoming data. When the current load is high, that lifespan shrinks. Streaming API on the other hands eliminates the Twitter server as a go between. Once the pathway is opened the tweets go directly to the device collecting them. However, tweets are only collected while the pathway is open, so to collect tweets over an extended period of time, the collector must maintain their own server that is constantly connected to the Internet and Twitter. As the tweet collection for this study was conducted from a personal laptop, this was not an option.

Second, many of the studies that analyze Twitter data, collect that data over multiple months and sometimes years. The current project had to adhere to deadlines, which limited the timespan over which tweet collection was possible (2+ months).

**Potential for Sentiment Analysis; Machine Learning Algorithms**

One of the limitations of the current sentiment analysis is that it relied on pre-defined lexicons. This is a limitation because the sentiment of the tweet can only be estimated when at least one word from the lexicon and one word from the tweet match up. The short length of tweets impedes this process but also the vernacular of the lexicon and the vernacular of the tweets don’t easily coincide. Tweets are short, often use slang and informal language, and are also prone to misspellings. In contrast most lexicons were developed to analyze the sentiment of full-length documents where somewhat formal written language is expected.
Sentiment analysis of tweets

This issue can be mitigated by modify existing lexicon to include even more slang or emoticons. Also, an all new lexicon can be constructed based on the language in tweets with help of Amazon Mechanical Turk (as in Mitchell et al., 2013). While not necessarily expensive, Amazon Mechanical Turk does cost money and so is limited to projects with funding. There are also Natural Language Processing techniques for fixing some of the misspellings in the tweets. A researcher with more experience with NLP might consider taking advantage of those.

Another option is employing machine learning algorithms such as, naïve Bayes classifiers, max entropy classifiers and random forest classifiers. Classifiers ‘learn’ patterns in the data given to them, so one must start with training data. Training data is data that the researcher is certain, through manual classification, is positive or negative. Training data is usually a list of positive tweets and a list of negative tweets. The classifier is given this training data and told to identify features and then observe the differences in patterns in those features when the tweets are positive and when the tweets are negative. Like in the case of the predefined lexicon the classifier might observe the feature ‘happy’ occurs most often in positive tweets. The difference is with lexicon positive words such as happy were decided \textit{a priori} to be positive while a classifier \textit{observes} that the word happy frequently occurs in positive tweets and deduces that if a tweet contains the word happy then it is more likely positive than a tweet that doesn’t. Also unlike lexicons, classifiers look for more features than just words, they can look for the joint occurrence of two words together. This can significantly improve accuracy, for example imagine the tweet “I am not happy”, the lexicon method would label the tweet positive because of the word ‘happy’. The classifier on the other hand might recognize that the combination of not and happy actually occurs frequently in negative tweets, which might overrule the probability of the tweet being positive because it contained the word happy. The primary caveat to classifiers is that they require very large training sets, because they can only develop information and patterns for features they’ve already seen.
INTERNATIONAL PERSPECTIVE

Traditional pen-and-paper data from developing countries is notoriously difficult to acquire. The reason for this varies. In some cases, the cost of accurate recordings is too great. In others, much of the population is difficult to reach. Finally, corruption within the government and other factors can impede the reliability of the data. Arguably, using Twitter as a data collection tool is best matched to combat corruption obstructions of data. In extremely rural countries or countries with undeveloped Internet infrastructure, Twitter data might not be ideal. Nevertheless, as Internet infrastructure grows around the world, twitter will gain greater relevance.

The methodology explored in this paper has the potential to be utilized in different regions of the world as those other countries develop and gain access to the social networking site. On a positive note, this may happen sooner than later as personal computers and tablets are becoming more and more affordable and projects are being planned to distribute them in countries with burgeoning economies. Additionally, Internet infrastructure is being funded and built in developing countries often faster than other forms of infrastructure. A private company in Rwanda has already begun laying a fiber optic network; two hundred kilometers around the capital city and plans for expansion to one thousand kilometers (Fripp, 2010). Additionally, a twenty-five dollar personal computer is also in the works to be available in developing countries, developed by the British nonprofit Raspberry Pi Foundation (Murphy, 2011). The spread of the Internet and personal devices might spread at a faster rate than one might expect.

Twitter and other social networking sites demonstrated their relevance in countries in Northern Africa and the Middle East during the Arab Spring. According to the Arab Social Media Report, at the end of June 2012 there were approximately 2,099,706 Twitter users in the Arab region and “Arabic is the fastest growing language in Twitter history” (http://www.arabsocialmediareport.com/Twitter/LineChart.aspx?&PriMenuID=18&CatID=25&mnu=Cat).
Sentiment analysis of tweets

Twitter and text sentiment analysis are both growing rapidly. Twitter is available in over 20 languages and used in nearly every country in the world, at least to some degree (www.twitter.com). Word affect lists are also becoming available in different languages. ANEW has one for Spanish, and a decent selection is already out there for Dutch and German. Additionally, machine-learning techniques don’t discriminate on the basis of language, which allows non-fluent researchers to study data from foreign countries.

Prospects are excellent for the researchers in foreign and/or developing countries and also for improving data collection in those same regions. With any luck, the methodology developed in this paper might contribute to that inevitable progression.

COLLABORATION

Behavioral Economics is an excellent example of what can be achieved by cooperation and sharing of knowledge between fields. The culmination of insights from economics and psychology has led to a much greater understanding of human behavior and how to understand and measure utility than either field has previously arrived at independently. Burgeoning research into harnessing Twitter data is ripe for collaboration. The current study is one of the first to explore the feasibility of measuring experienced utility by way of Twitter, and the implications for valuing public goods and environmental externalities thereafter. Computer and information scientists have thus far conducted the bulk of Twitter research, however, as evidenced by this study it stands to benefit from economic and international perspectives. In the same vein, as discussed, I did not have the computer science and natural language processing background to perform as thorough a sentiment analysis as I might have liked. Both a theoretical perspective and a technical perspective are pertinent to the progression of Twitter research.
CONCLUSION

In this paper we investigated whether by analyzing the language in tweets for positive and negative affect we could remotely monitor the experienced utility of cities. This investigation was motivated in two principal ways.

First, we were unconvinced of conventional economic utility models ability to realistically reflect human behavior and consequently skeptical of relying on them to determine what factors improve quality-of-life and well-being. As a planet we have reached a predicament. The lifestyle developed primarily in the western hemisphere over the last few decades puts an unsustainable strain on the Earth’s resources. The economic theory that insists on an ever-growing GDP is partially to blame for the situation in which we have found ourselves. Many developing countries are at a pivotal moment in their development and are deciding how to approach improving quality-of-life within their borders. Faced with this challenge, it is important to reflect on and re-evaluate the paths we took as a society in the past. Research in behavioral economics and psychology has laid the bricks to consider alternatives to the conventional economic utility theory.

One viable alternative is experienced utility, but an agreed upon method for measuring experienced utility has not been reached. The current study assessed the validity of measuring experienced utility at the city level with collection and sentiment analysis of tweets. Twitter has the benefit being cheap, widespread, and with an extremely current data set that can be accessed from anywhere with an Internet connection. The current study analyzed tweets from cities across the United States measured against US Census data. We found that there was sufficient variation in the tweet affect construct. We also found that it correlated in the correct direction with variables for which we had a priori predictions and considered important factors for quality-of-life. The same collection and analysis can easily be expanded to other countries where Twitter is used and independent variables similar to those accessed through the US Census Bureau can be collected. From there, cross country analyses comparing models from different countries can be conducted. The primary limiting factor is the level of
Sentiment analysis of tweets

usage of Twitter in some countries. However both Internet infrastructure and Twitter are constantly expanding so this limitation is expected to shrink in the near future.

While the primary goal was to provide evidence for the validity of measuring experienced utility with tweets scored for affect, we also found interesting relationships between the independent variables and the tweet affect score in the models. For example, the percentage of the population that commutes to work by bike was found to be consistently positive and significant while a negative and significant relationship was found with median earnings of the population 25 years and older. The former finding says cities with more cyclists are on average more positive. This suggestion is particularly intriguing, because its assigns some value to a variable otherwise difficult to abstract a value for. Measuring experienced utility opens the doors to novel findings like bicycling.

However it is important to recall that experienced utility is a theoretical construct that relies heavily on the hedonic theory of happiness. We really don’t have any scientific evidence that hedonism is the true path to happiness. There is no scientific evidence to say we should want to maximize happiness, positive/negative affect, or global life evaluations. Rather as a society we are constantly evaluating and re-evaluating our goals and priorities. Different approaches, research, and measurements help us do this. Measuring experienced utility with tweets adds to that toolbox.
Sentiment analysis of tweets

REFERENCES


Sentiment analysis of tweets


55
Sentiment analysis of tweets


Sentiment analysis of tweets


