

# Estimating Consumer Demand for Information Attributes Using Twitter Data\*

Ben Smith<sup>†1</sup> and Jadrian Wooten<sup>‡2</sup>

<sup>1</sup>College of Business Administration, University of Nebraska at Omaha

<sup>2</sup>Department of Economics, The Pennsylvania State University

2015

## Abstract

In 2006, a new social web service was launched: Twitter. That service was centered around the idea that each user could follow other users, but there was no requirement for reciprocation. This lack of reciprocation between creators of content, and those who consume the content, exhibits an advantageous property for social scientists: you directly observe the demand curve (if a set of criteria is satisfied). This means that anytime information is in demand and the consumer pays for it with their time, Twitter data can be used to estimate the information attributes in demand by the consumers. We have developed a method, as well as software, to exploit this environment. Our method can be used by consumer economic researchers, environmental researchers, marketing departments in industry, government agencies, and many others.

*JEL classification:* C81, C88, D12

*Keywords:* data collection, regular expression, twitter, demand estimation, social media

---

\*The authors thank Roy Fletcher as well as the other conference participants at the 2013 Academy of Economics and Finance Conference. We also thank Philip Wandschneider, Robert Rosenman, Jill McCluskey, Gnel Gabrielyan, Jared Woolstenhulme and Andrew Cassey.

<sup>†</sup>Corresponding Author, Assistant Professor, College of Business Administration, University of Nebraska at Omaha. Mammel Hall, Suite 332, 6708 Pine Street, Omaha, Nebraska 68182. E-Mail: bosmith@unomaha.edu. Phone: 402.554.2803. Fax: 402.554.2853.

<sup>‡</sup>Lecturer, Department of Economics, The Pennsylvania State University. 315 Kern Graduate Building, University Park, Pennsylvania 16802, E-Mail: jjw27@psu.edu. Phone: 814.865.7352.

# 1 Introduction

The past few years have been characterized by marked increase in social network usage (Duggan and Brenner, 2013). People spend large portions of their life online discussing their activities, the things they like and other aspects of their desires. As of 2015, one of the most popular networks is Twitter: enjoying 300 million active users as well the creation of 500 million tweets (the messages created by users) per day (Twitter, 2015).

Consumer preferences are trapped in paragraphs of text on various social sites. This is certainly not a revelation. However, using language analysis (a process underused in the economics field), we can transform these attributes from language into data that can be analyzed using statistical methods. Our method and software allow the analyst to extract data from Twitter and use language analysis to convert attributes to numerical data. This technique creates large, timely, and inexpensive datasets.

The outline of the paper is as follows: (1) Discuss the potential demand exhibiting characteristics of Twitter when information is in demand, (2) Consider the potential bias in this approach and compare it to other methods of data collection, (3) Discuss the process of collecting and analyzing Twitter content, (4) Show an example using pundit predictions from the 2013 Super Bowl, and (5) Define a list of best practices.

## 2 Twitter: A (Potential) Demand Curve Waiting for Extraction

Twitter is a social media site where individuals “follow” the messages (or “tweets”) of other individuals. When an individual tweets a message, that message appears in the “stream” of any individual following them. However, the reverse is not true: when a follower tweets a message, the people they follow will not see the message.

While it is free to follow any individual, it is not costless. As an individual follows more people, their stream becomes increasingly crowded. If the individual follows “too many” people, the content they care about becomes buried within content they do not care about (i.e. congestion costs). It follows that there exists a limit (which may be different for each individual) to the number of users a person can follow efficiently, and this limit can be described in terms of tweets over some time interval.

A content creator, however, experiences no cost when they gain a new follower. They are not notified in any way<sup>1</sup> and it does not change the experience of producing a tweet.

Because there is no cost to an additional subscriber. A tweet produced by a supplier is a classic public good, however the supply of producers (compared to the supply of tweets) is not. The motivation of suppliers to enter the Twitter content creation market will be different with every social science question.

In some cases, this concern can likely be satisfied. Media producers may view Twitter as an inexpensive promotional vehicle for their other outlets (where they usually generate revenue through advertising). The cost of alternative advertising in 2013 is approximately \$1.44 (Shalvey, 2013) per click-through (Google AdWords); that means as long as the media advertiser considers maintaining a Twitter presence to be cheaper than the equivalent amount of engagement through alternative means, they will maintain a Twitter presence. Situations where the producer can maintain even

---

<sup>1</sup>One exception is if the user has a private account. Private accounts make up about 2% of Twitter’s user base (Wagemakers, 2012) and are not reflected in any study using Twitter as the data is restricted and not published publicly.

moderately large followings will satisfy this condition. Once they decide to engage in Twitter activity, these producers wish to supply as many people as possible on Twitter as it increases their overall advertising for the same fixed cost.

However, reputation maximizing accounts may not satisfy this condition. Consider the customer support account for a cable company: the goal of the technicians managing the Twitter account is to solve the problems presented by users such that the issue does not “go viral” and become a public relations problem. Companies where this is a low probability event may not even have a Twitter account (as it may cost more to manage than the expected cost of a public relations incident). For these reasons, a study based on the tweets created by support accounts would likely not be able to assume the supply of content creating accounts is fixed.

Assuming the supply conditions can be satisfied for the given question, you are observing the demand curve directly for information on Twitter. From this we can make a general statement - quantity demanded (followers) is a function of explanatory and control variables:

$$\mathbf{F} = D(\mathbf{X}; \mathbf{C}) + \varepsilon \tag{1}$$

Where  $\mathbf{F}$  is a vector of followers,  $\mathbf{X}$  represents explanatory variables and  $\mathbf{C}$  is some set of control variables. Total followers of an account are available from the Twitter streaming application programming interface (API) along with many control variables such as age of the account (accounting for discovery) and the number of favorite tweets (accounting for engagement). Explanatory variables are generally extracted from the tweet itself using language analysis.

Others have used Twitter data in the computational sciences. Kwak et al. (2010) mapped the network relationships for distributional characteristics. Dong et al. (2010) examined the relationship between links in tweets with search ranking and Bollen et al. (2011) used the ‘mood’ on Twitter to predict the stock market. However, to our knowledge no one has used Twitter to explore consumer preferences.

Many news outlets have used Twitter information but in a ‘crude’ fashion. During the 2012 presidential debates, all three major networks used the trending keywords (simply counting the number of times a word is tweeted, a metric provided by Twitter) as a benchmark of the popularity of certain topics. Academics have used more sophisticated methods of analyzing hashtags (Lindgren and Lundström, 2011), but such research is rare.

Many economic studies have focused on consumer demand attributes from consumer attitudes toward organic food to privacy (e.g. Huang (1996); Keen et al. (2004); Murray (1991); Phelps et al. (2000)). The methods used in such papers (e.g. surveys, experiments and access to protected data) often include expensive datasets (either by method of collection or fees), old data, or at a minimum, small datasets. Our method, outlined below, involves large datasets which are inexpensive to collect and are timely in nature. This is not to in any way minimize the importance of other methods outlined. However, what we supply is an additional tool in the analyst’s tool chest.

For instance, those interested in political dynamics could study the attributes of news presentations that consumers find most appealing. If the analyst is working for a non-profit they might estimate the most effective way to deliver their message. Academics might estimate interest in public goods.

Not only can data be collected about consumer preferences, but the degree of those preferences (the ‘score’) can be analyzed using appropriate language analysis.

## 2.1 Potential Bias of Data and Comparison to Other Methods

Like any method in this field, the data on Twitter is not perfectly representative of the U.S. population as a whole. Because people self-select into the use of Twitter, there is an inherent bias. However, it matches some demographics nicely. According to Pew (Duggan and Brenner, 2013), the Twitter population represents national income and education demographics fairly well, but is marginally younger (the 18-49 age group is overrepresented and the over 65+ is underrepresented) and slightly more urban than the general population.

Additionally, there are probably unobservable characteristics that could potentially be different from those on Twitter compared to those off of Twitter (due to the self-selection issue). For instance, a reasonable argument could be made that those on Twitter have less concern about privacy (given that they are on a network, that by its nature is public) and are perhaps more technical (given the network used to be centered around the use of text messages). Anecdotally, it appears that media personalities, and those who are interested in the media, are overrepresented.

However, Twitter also avoids some types of bias. Some common problems associated with survey and experimental data include the ‘observer-expectancy effect’ (Rosenthal, 1964), ‘demand characteristics’ (Nichols and Maner, 2008) and the ‘Hawthorne effect’ (McCarney et al., 2007). In the case of the observer-expectancy effect, the experimenter/surveyor unconsciously sends signals to the participant that influences their response. The demand characteristics are characterized by the participant either wanting to please or displease the experimenter/surveyor. As a result, participants often attempt to choose the ‘right’ answer (for their desired outcome) instead of the honest answer. Finally, the Hawthorne effect is a common tendency of participants to perform better when they know they are being observed. Some, however, have argued that the Hawthorne effect is simply a variant of the demand effect (Adair, 1984). We will colloquially refer to all of these effects as ‘observer effects.’

Because Twitter is an active real-world marketplace, and the users are unaware data collection is occurring, the observer effects are likely less concerning. Additionally, the selection issues are fundamentally different. With surveys and experiments, in most cases the analyst, for practical reasons, must obtain volunteers who know they are being observed for data collection from a geographically un-diverse region.

Whether the potential bias of Twitter data is of concern (or at least less concerning than other methods) is dependent on the question being asked. There are certainly situations where the bias would be so great that it would render the use of the network unadvisable. For instance, those studying the potential technical features desired of a program targeted at senior citizens would likely generate results biased by technical abilities of those on versus off Twitter. Alternatively, studying preferences for entertainment through Twitter may be substantially better than other methods due to the avoidance of the observer effects.

## 3 Collecting and Analyzing Data

### 3.1 Extracting Data

Twitter allows any internet user to programmatically “listen” to a set of words (this is referred to as the “Streaming API”). If a specific set of words exists in any given tweet, the tweet is sent to a running computer program in real time (software must be running on a computer connected to the Internet to collect the tweet).

We have developed software that can collect tweets on any set of watch words and save them to a SQLite database. SQLite is a high performance format that can be opened by many other programs. This data includes all of the information about the tweet and content creator that Twitter provides including: followers, the user biography, the tweet text, verified status, status update count and much more<sup>2</sup>.

Please note, to use our software you need not be a programmer, you simply set your Twitter developer credentials (Twitter, 2012) and specify a plain text file (specifying the words you wish to watch) via the command line.

### 3.2 Analyzing Data

Once the data is collected one needs to analyze it. Our software uses a technique called regular expression where a large number of phrases can be generalized such that variations can match a single criterion (or expression).

For example, suppose we are interested in how different ways of describing particular energy technologies (e.g. word choice) lead to differing levels of popularity. Under this scenario, one of our regular expressions might be something like<sup>3</sup>:

```
\b((clean|smokeless)[\s]coal)(?:?!(\b((not)|(won[']t))\b)).)*\b(fantastic|↔  
awesome|great)\b
```

Which would match any sentence matching a particular opinion about clean coal while excluding sentences meaning the exact opposite. Using a spreadsheet program, the analyst creates a large set of these expressions in one column followed by a score (or value) to assign the tweet if the expression is matched. The analyst can have as many rows of regular expression (with scoring) as they desire (as well an infinite number of scoring levels).

Once the analyst has their set of expressions in comma-separated values (CSV) format, as well as the SQLite database of collected tweets from the collector (Section 3.1), they can use our analyzer software to extract the scores from the tweets (you can also analyze the user biography). Using our software, the analyst simply specifies the SQLite database containing the tweet data as well as the CSV file containing the regular expressions (one can also run the analyzer software multiple times if you have multiple, distinct, scoring criteria).

When instructed, the analyzer loops over all of the specified regular expression patterns over every tweet in the SQLite database, when it finds a match it saves the result in a separate table in the database. Given a modestly long list of regular expressions, this process can involve thousands of iterations for each tweet processed. This can potentially take a large amount of time, so the program keeps track of the tweets processed such that it can recover from a crash, or restart, and indicates to the user periodically of its progress.

Finally, once the process analyzer is complete, the user can access the processed data directly from a statistical package that reads SQLite databases or can instruct the analyzer to export a CSV.

---

<sup>2</sup>Twitter allows an application to listen to up to 400 words at a time, where the collecting application receives any message containing one of those words (Twitter, 2013). Our software (instructions in Appendices A and B) saves these messages in SQLite. SQLite is an open source database (<http://www.sqlite.org/>) that can be opened by many applications including R (R Core Team, 2015).

<sup>3</sup>We are using a standard full implementation of regular expression – a standard pattern matching syntax (for documentation see <http://regular-expressions.info/>). Therefore, there is no restriction on the complexity of the regular expressions.

## 4 Demand for Pundits: The 2013 Super Bowl

As a demonstration of the method, we present the following example. The week prior to the 2013 Super Bowl we collected predictions from both amateur and professional pundits about the game<sup>4</sup>.

We theorize that consumers demand both accuracy and confidence (as in using confident, or strong, words) from sports pundits due the psychological costs of uncertainty (Osuna, 1985). For the purposes of this example, we defined a pundit being accurate when they correctly predicting the winner of the 2013 Super Bowl. A pundit is considered confident when they use strong words when making that prediction (as defined by Chklovski and Pantel (2004), who ranked the relative strength – or confidence – of words)<sup>5</sup>.

Pundits are likely attempting to advertise their outside activities (such as being on television or on a website), therefore it is in their interest to have as many followers as possible. The lowest cost alternative advertising is likely Google AdWords which is about \$1.44 per engagement. It seems relatively safe to assume to that the minimum number of followers is low enough that we are seeing a nearly full spectrum of outcomes.

We are interested in the response of most consumers, so the Twitter demographics are mostly advantageous to our question. Additionally, due to vast differences in geographic responses (we are examining sports), we need a sample from the entire country and not a small geographic region. For these reasons, we believe the benefits of this method outweigh the negatives given our particular question.

Using the collector (Section 3.1), we created a text file containing the two teams’ names, cities and nicknames. This resulted in collecting any tweet containing one of those words. Because you have to specify at least one of the team names (or cities/nicknames) to specify an outcome, this is sufficient to collecting any prediction. Once the game has started, we stopped the collector.

Using the analyzer (Section 3.2), we used a table of regular expressions, in this case expressions that match particular predictions at differing levels of confidence (the “score” referred to in Section 3.2) to turn the text of the tweets into numerical data. We further used the analyzer to determine if the tweeter claimed sports expertise in their biography section and if they were correct about the outcome of the game. These results (those claiming sports expertise and making predictions about the game with varying levels of success and confidence) were then exported as a CSV using the analyzer.

### 4.1 Results

#### 4.1.1 Summary Statistics

Using the data from the analyzer, we observe the summary statistics in Table 1.

### 4.2 Model

We propose the following model for sport punditry:

---

<sup>4</sup>This is a small example using a limited dataset. The purpose here is to be illustrative of the method, not conclusive about pundits. For our example, we define a ‘pundit’ as any person who either claims expertise (amateur) or is hired to give their opinion in the media (professional) on a particular subject.

<sup>5</sup>For the purposes of this study, we used a dummy variable where 1 indicates a confident phrase and 0 indicates a less than confident phrase. As the work by Chklovski and Pantel (2004) was ordinal in nature, we used the more confident (strong) half of the range provided as an indication of a confident phrase.

Table 1: 2013 Super Bowl Summary Statistics (1159 Observations)

	Coefficient	Mean	Std. Dev.
Confidence	$C$	.258	.438
Accuracy	$A$	.528	.499
Tweets per Year	$M$	5597.207	9135.366
Account Age	$G$	2.769	1.248
# of Days before Game	$B$	2.897	2.294
Tweeter Favorites Count	$Fv$	75.311	245.901
Followers	$F$	2747.921	9457.935

$$\begin{aligned}
 \text{Log}(F_i) = & \alpha + \beta_1 C_i + \beta_2 A_i + \beta_3 M_i \\
 & + \beta_4 M_i^2 + \sum_{k=1}^j \gamma_k X_{ik} + \beta_5 V_i + \varepsilon_i
 \end{aligned}
 \tag{2}$$

Where  $F_i$  is the number of followers (quantity demanded),  $C_i$  is the subjective confidence (0 or 1),  $A_i$  is accuracy (0 or 1),  $M_i$  is the number of messages per year,  $X_{ik}$  is a set of control variables, and  $V_i$  is if the account is verified or not (1 or 0 – are they celebrity/professional pundit).

This functional form appears to be relatively appropriate and it is the best behaving common functional form we have tested<sup>6</sup>. Further, we expected a log-linear model to be the proper fit. Assuming that there is a massy center of subscriber (Twitter follower) preferences, one would expect that the supplier would get increasing returns as they better match common preferences.

### 4.2.1 Regression Results

Using a standard iterated GMM approach, we find the following coefficient values ( $\beta_*$  indicates significant at 95%):

$$\begin{aligned}
 \text{Log}(F_i) = & 3.554 + .152C_i + .306_*A_i + .000_*M_i - .000_*M_i^2 \\
 & + .765_*G_i - .081_*G_i^2 - .006B_i + .003_*Fv_i + 1.647_*V_i + \varepsilon_i
 \end{aligned}
 \tag{3}$$

$G_i$  is the account age in years (accounting for discovery),  $B_i$  is the number of days before the game the prediction was made and  $Fv_i$  is the favorite count (the number of items the content creator has favorited – a measure of engagement). The adjusted  $R^2 = 0.545$ . While both log-followers and residuals appear to be normal (using stem-and-leaf and histogram), a Shapiro-Wilk test rejects the null, so bootstrapping (at 10,000 replications) is used to account for any distribution issues and jackknifing is used so the results can be replicated (Table 2).

### 4.3 Discussion

In our simple example above, we are able to use tweets to show that consumers likely demand accuracy ( $A$ ) and confidence ( $C$ ) from their pundits, but we can only say that with 90% statistical

---

<sup>6</sup>Adding higher order terms to the tweets per year ( $M_i$ ), account age ( $G_i$ ), days before game ( $B_i$ ), favorites ( $Fv_i$ ), as well as combinations, did not result in substantive changes in the coefficient values of interest (confidence ( $C_i$ ) and accuracy ( $A_i$ )) or substantive improvements of fit.

We estimate  $\lambda = -0.06$ , which is close to zero. However, we also re-estimated our model using  $\lambda = -0.06$  for the curious reader in Appendix C. Our data is available at <http://goo.gl/Isi28>.

Table 2: 2013 Super Bowl GMM Bootstrapped & Jackknifed Standard Errors and P Values

Coefficient	Bootstrapped Std. Err. (10,000 Replications)	$P >  z $	Jackknifed Std. Err.	$P >  t $
$C$	.086	.076	.087	.080
$A$	.074	.000	.074	.000
$M$	.000	.000	.000	.000
$M^2$	.000	.000	.000	.000
$G$	.148	.000	.149	.000
$G^2$	.028	.004	.028	.004
$B$	.015	.701	.015	.703
$Fv$	.000	.000	.000	.000
$V$	.207	.000	.219	.000

$C_i$  is confidence,  $A_i$  is accuracy, both range from  $[0, 1]$ .  $M_i$  is ‘tweets’ per year and  $V_i$  is if the account is verified. Control variables include  $G_i$  (account age),  $B_i$  (before game time) and  $Fv_i$  (favorites count)

confidence. This is inline with the psychological literature and our hypothesis. It is also shown that, unsurprisingly, professional pundits have larger numbers of followers than amateurs ( $V$ ) – they accumulate followers from outside activities (like being on TV). Both the message creation rate ( $M$ ) and account age ( $G$ ) have positive but diminishing effects (in percent terms). On a network like Twitter, allowing time for people to discover you is clearly important given the coefficient sizes. Perhaps one of the most surprising results is that while both the message rate coefficients are statistically different from zero, neither is large enough for us to care about them in an economic sense. However, this can be explained by the two opposing forces that contribute to the message rate coefficient: cost to existing followers (negative – increasing tweets increases the opportunity costs) and discovery by non-followers (positive – discovery increases the more the user tweets).

Using the method outlined in this paper with the tools we make available as well as a statistical package, we collect predictions from Twitter, analyze them, and then perform statistical tests. This basic procedure can be replicated by any social scientist given that there is no special skills or knowledge required to use our tools.

## 5 Best Practices

In this section we outline a set of best practices that should be followed, if possible, when using Twitter data to estimate consumer preferences.

### 5.1 Defined Demand Model Condition

First, prior to any data collection or analysis, the analyst must be clear on what is in demand and how they intend to measure the properties of interest using language analysis. Because the analyst supplies a list of ‘watch words’ to the collector, if the model is unclear to the analyst at time of collection, they might watch the wrong words.

### 5.2 Supply Side Conditions

For the estimation of demand to be valid, the supply side must be follower maximizing. In effect, the desired number of followers on the supply side must be infinite. When considering a research question, the analyst should ask the following questions:

- Is the goal of the supplier of tweets to maximize exposure?

- Is the alternative to supplying tweets expensive enough that it is reasonable to assume that a large representative sample of suppliers have entered into the Twitter market?

### 5.3 Demand Conditions

If the supply side conditions are satisfied, then the analyst is observing a demand curve. However, that may not mean the Twitter demand curve is representative of anything outside of Twitter. For the analyst to draw broader conclusions, the following must be satisfied:

- Is the Twitter demographics representative of the underlying population of interest?
- Is there no/few reasons to believe that the participation in a social network, that is inherently public, would result in a biased sample?

### 5.4 Comparison to Other Methods

Finally, one should consider the bias compared to other methods.

- Is the observer effects likely large or worse than the demographic or supply side concerns of Twitter?
- Is the demographic concerns or other experiment/survey related selection bias worse than the bias in Twitter data (for a particular question)?
- Is performing a survey/experiment cost prohibitive?

When possible, it is also helpful to perform tests on the collected data (such as testing if the tweet flow is correlated to the number of followers).

## 6 Conclusion

In this article, we outline the subscriber market on Twitter, explaining that, when it comes to demand for information, Twitter can exhibit a demand curve when the supplier wishes to supply to an infinite number of people. Further, we discuss the use of our custom built tools that allow any social scientist to collect and analyze data from Twitter. We demonstrate an example involving the 2013 Super Bowl using only the tools we make available to the analyst and a statistics package. Finally, we define a set of best practices when using Twitter to explore consumer preferences.

We've make our tools' source code available for use and modification. However, we have also published our software to a peer-reviewed repository so that it is easy to install and update (Mac-Ports) and created precompiled binaries for Microsoft Windows. This method makes it possible for academics, government, and industry alike to start exploring consumer preferences inexpensively and rapidly.

## References

Adair, J. G. (1984). The Hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology*, 69(2):334, doi:10.1037/0021-9010.69.2.334.

- Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, doi:10.1016/j.jocs.2010.12.007.
- Chklovski, T. and Pantel, P. (2004). Verbocean: Mining the web for fine-grained semantic verb relations. *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 4:33–40.
- Dong, A., Zhang, R., Kolari, P., Bai, J., Diaz, F., Chang, Y., Zheng, Z., and Zha, H. (2010). Time is of the essence: improving recency ranking using twitter data. *ACM Press*, pages 331–340, doi:10.1145/1772690.1772725.
- Duggan, M. and Brenner, J. (2013). The demographics of social media users – 2012. [http://www.pewinternet.org/files/old-media/Files/Reports/2013/PIP\\_SocialMediaUsers.pdf](http://www.pewinternet.org/files/old-media/Files/Reports/2013/PIP_SocialMediaUsers.pdf). Accessed: 2015-06-22.
- Huang, C. L. (1996). Consumer preferences and attitudes towards organically grown produce. *European Review of Agricultural Economics*, 23(3):331–342, doi:10.1093/erae/23.3.331.
- Keen, C., Wetzels, M., de Ruyter, K., and Feinberg, R. (2004). E-tailers versus retailers: Which factors determine consumer preferences. *Journal of Business Research*, 57(7):685–695, doi:10.1016/S0148-2963(02)00360-0.
- Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media? *ACM Press*, pages 591–600, doi:10.1145/1772690.1772751.
- Lindgren, S. and Lundström, R. (2011). Pirate culture and hacktivist mobilization: The cultural and social protocols of #WikiLeaks on Twitter. *New Media & Society*, 13(6):999–1018, doi:10.1177/1461444811414833.
- McCarney, R., Warner, J., Iliffe, S., van Haselen, R., Griffin, M., and Fisher, P. (2007). The Hawthorne Effect: a randomized, controlled trial. *BMC Medical Research Methodology*, 7(30), doi:10.1186/1471-2288-7-30.
- Murray, K. B. (1991). A test of services marketing theory: consumer information acquisition activities. *Journal of Marketing*, 55:10–25, doi:10.2307/1252200.
- Nichols, A. L. and Maner, J. K. (2008). The good-subject effect: Investigating participant demand characteristics. *The Journal of General Psychology*, 135(2):151–166, doi:10.3200/GENP.135.2.151-166.
- Osuna, E. E. (1985). The psychological cost of waiting. *Journal of Mathematical Psychology*, 29(1):82–105, doi:10.1016/0022-2496(85)90020-3.
- Phelps, J., Nowak, G., and Ferrell, E. (2000). Privacy concerns and consumer willingness to provide personal information. *Journal of Public Policy & Marketing*, 19(1):27–41, doi:10.1509/jppm.19.1.27.16941.
- R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rosenthal, R. (1964). The effect of the experimenter on the results of psychological research. *Bulletin of the Maritime Psychological Association*, 13(1):1–39.

- Sedgewick, R. (2014). Using a command line interface. <http://www.cs.princeton.edu/courses/archive/fall14/cos126/precepts/CommandPromptTutorial.pdf>. Accessed: 2015-06-24.
- Shalvey, K. (2013). Google per-click ad rate seen up first time in year. <http://news.investors.com/technology/032813-649743-google-ad-rate-seen-increasing-in-first-quarter.htm>. Accessed: 2015-06-22.
- Twitter (2012). Tokens from dev.twitter.com. <https://dev.twitter.com/docs/auth/tokens-devtwittercom>. Accessed: 2015-06-22.
- Twitter (2013). Post statuses/filter. <https://dev.twitter.com/docs/api/1.1/post/statuses/filter>. Accessed: 2015-06-22.
- Twitter (2015). Twitter usage / company facts. <https://about.twitter.com/company>. Accessed: 2015-06-22.
- Wagemakers, A. (2012). The last 100 million twitter accounts. <http://www.twopblog.com/2012/05/last-100-million-twitter-accounts.html>. Accessed: 2015-06-22.

# Appendices

## A Installing Our Tools

Appendices A and B provide a brief overview the installation and use of our software tools. However, if reader is unclear on any step or wish to extend one of the tools, the authors are available via e-mail at [twitter@bensresearch.com](mailto:twitter@bensresearch.com).

Regardless of your operating system, you will use the program by using the command prompt or terminal. If you have never used a command line tool, we recommend that you read an online primer on the subject. For Mac and Windows users, an excellent primer is available from Sedgewick (2014).

A more advanced user can acquire our source code from <http://goo.gl/LXieoo>. However, we have also made the code available as ‘ports’ on the Mac and precompiled binaries on Windows. If you are inexperienced with underlying programming languages, you will want to obtain our software either as Mac ports or as precompiled binaries.

For Mac users, one first must install MacPorts. Thankfully, installing MacPorts is fully explained on their main website located at <http://www.macports.org/>. Once MacPorts is installed, you can install our software by simply typing the following into the terminal:

```
sudo port -v selfupdate
sudo port install TwitterDemandCollector
sudo port install TwitterDemandAnalyzer
```

For Windows users, you can download our precompiled binaries from the following locations:

- TwitterDemandCollector: <http://downloads.bensresearch.com/TwitterDemandCollector.zip>
- TwitterDemandAnalyzer: <http://downloads.bensresearch.com/TwitterDemandAnalyzer.zip>

Both downloads are archives with a single executable file. *When you are on Windows, you must navigate the command prompt to the directory where you have download the two precompiled binaries* (as described by Sedgewick (2014)). For the following instructions, we will assume you’ve located both binaries in a single folder and you’ve navigated to that folder in the command prompt.

## B Using our Tools

In the main body of the manuscript we concentrate on conditions when it would be advantageous to use Twitter data for research. In this appendix, we focus on the logistics of using our tools. Once you have followed the advice we have laid out in Section 5, you should have some idea of what set of ‘words’ you need to collect and the regular expressions you will use to analyze the tweets. These two lists map to two files that you must create. We will assume the words file is named ‘words.txt’ and shown in Figure 1.

In this example, we will be collecting any tweet that contains at least one of the following (case insensitive) words: market, IBM, MSFT, AAPL, S&P. In this case, it is clear we are doing some sort of analysis related to a subset of stocks. However, this list of words could be anything.

### B.1 Collecting Data

To collect data from Twitter, we must first create a Twitter application. Thankfully, you will only need to do this once. From <https://apps.twitter.com/>, follow the onscreen instructions to create



Figure 1: The collection ‘words’ file shown in a plain text editor.

an application. The name of the application, description and web address do not matter for our purposes. This is an application that only you, the researcher, will access. This is simply a way to create tokens that the collector will use. Once you have created the application, follow the instructions outlined in Twitter (2012) to create application tokens.

At this point, your Twitter application should have all of the ingredients necessary to start collecting data. Now, you simply need to inform the collector of this information. From the ‘Keys and Tokens’ sub-tab on the Twitter Application Management site, you will find information similar to Figure 2.

To enter this information into the collector, type something similar to the following – replacing your own keys, secrets, and tokens in the appropriate fields:

```
TwitterDemandCollector --key="oYeU0iDtqP1ng7khfQy8qVrxa" --key_secret="↵  
eF60qBGaAZEB0cfAJtRXRsk28PMaSSb4rPJtDjE2IqW7sO3P9t" --token="3017325254-↵  
XG6W600LXNDwZAsLewn0BwVuJqNFgm3fq8iJGrh" --token_secret="↵  
C7APbnAAxldAAZsfanyXAFyLPdITOjlo4KxuvV2OnuDd"
```

Once you have hit ‘enter,’ the keys, tokens and secrets are saved to a configuration file. The collector will use them for all conversations with Twitter.

The next step is to load the set of words. Assuming that the file ‘words.txt’ is in the same directory as the collector application, the following command will load the list of words into the application:

```
TwitterDemandCollector -f "words.txt"
```

Typing the following command:

```
TwitterDemandCollector --printwords
```

will verify that the words have been imported into the program. The application is now set to collect data from the Twitter API. To begin collecting data, type the following command:

```
TwitterDemandCollector --run
```

While the collector is running, you have a number of tools to monitor the collection process (Figure 3). At any given time, you can retrieve the total number of tweets collected (count), the time of the most recent tweet (status), print the most recent tweets collected (print), or exit. Once you exit the collector (by typing ‘exit’), you will find a database file named ‘tweets.db’ within the same directory.

## Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	oYeU0iDtqP1ng7khfQy8qVrxa	← 1
Consumer Secret (API Secret)	eF60qBGaAZEB0cfAJtRXRsk28PMaSSb4rPJtDJE2lqW7sO3P9t	← 2
Access Level	Read and write (modify app permissions)	
Owner	Econ_Help	
Owner ID	3017325254	

### Application Actions

Regenerate Consumer Key and Secret    Change App Permissions

## Your Access Token

This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access Token	3017325254-XG6W600LXNDwZAsLewn0BwWuJqNFgm3fq8iJGrh	← 3
Access Token Secret	C7APbnAAxldAAZsfanyXAFyLPdlTOjlo4KxuvV2OnuDd	← 4
Access Level	Read and write	
Owner	Econ_Help	
Owner ID	3017325254	

Figure 2: The collector authenticates using four pieces of information: the key, the key secret, the token and the token secret. These keys and tokens are for demonstration purposes only and have been revoked. You will need to create your own keys and tokens.

## B.2 Analyzing Data

Having completed the collection process, a database file name 'tweets.db' will be within the active directory. The first step in analyzing the data is to import the database into the 'clean' format:

```
TwitterDemandAnalyzer --tweets="tweets.db" --clean="clean.db" --import
```

There are slight structural differences between the two database formats. When the collector is running, the database structure is designed to accommodate rapid saving, while the analyzer's clean format is designed to accommodate rapid retrieval.

Now is the time to analyze the large collection of tweets using regular expression. Suppose the regular expressions and scores are shown in Figure 4 with the file name 'regex.csv'.

As stated in the main body of the manuscript, regular expressions can have any level of complexity and the scores only need to have meaning to the researcher. Loading and running the regular expressions is accomplished by typing the following command:

```
TwitterDemandAnalyzer --clean="clean.db" --regex="regex.csv" --find -- ←  
processedtable="result"
```

In this case, each tweet that matches a particular regular expression is saved to a new table as specified by *processedtable*. Be forewarned, this processing can take a very long time. If you wish to run differing sets of regular expressions, you simply need to specify a different source file (*regex*) and destination table (*processedtable*).

```

http://t.co/XNi92jffwj via @Stock_Market4U

#Stock Market
:)
2015-06-24 13:22:46+00:00: | Prince and DeVos Families at Intersection of Radical Free Market Privatizers & Religious Right http://t.co/YE9TaiQJc #faithforward #1u
2015-06-24 13:22:45+00:00: RT @ibmmobile: IBM MobileFirst can help strengthen #BYOD security. Learn more: http://t.co/o0oP2iLTgB #IBMMobile http://t.co/RoUK7Sr5xV
2015-06-24 13:22:45+00:00: InFocus to Focus on Budget and Premium Smartphone Segments in India: "We have mapped the market in terms of se... http://t.co/0EM0hniXlA

You can do the following actions:
* exit
* count
* countuser
* print
* printuser
* status
* locations
>

```

Figure 3: During the collection process, you monitor the progress of the collection application by typing ‘count’, ‘print’ or ‘status’ and pressing enter. When you are ready to stop collecting data, typing ‘exit’ will close the application.

Once the processing is complete, you can export the results to a CSV by typing the following command:

```

TwitterDemandAnalyzer --clean="clean.db" --processedtable="result" --export="out.csv"

```

A sample of the output from this command is shown in Figure 5. You will note that we specify both an output file (‘out.csv’) and the processed table (‘result’). If you had created multiple processed tables, you would specify the name of the other table(s) to export those results. It is of note that you can combine multiple output files using the unique tweet identifier ‘twitterid’.

## C Box-Cox Transformed Dependent Variable

For interpretation reasons we used the logged dependent variable form in the main body of the text. However, we also estimated a Box-Cox fit:  $\lambda = -0.06$ . Using the estimated  $\lambda$ , the iterated GMM results in a .114 coefficient value for confidence ( $C_i$ ) and .233 for accuracy ( $A_i$ ). Other parameters experienced a scaling effect as well, but no parameter switched significance or was otherwise interesting. Other variable transforms attempted did not substantively change the results.

Using the coefficient values calculated from the Box-Cox transformed regression, Table 3 shows the percent gain in followers associated with a prediction being 50% or 100% confident or accurate (Table 3).

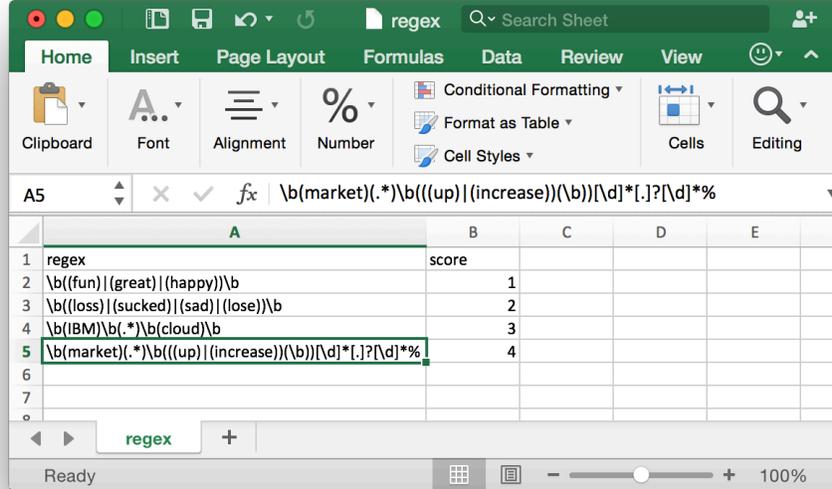


Figure 4: The regular expression CSV file as shown by Microsoft Excel. The regular expressions and scores can be entered in any spreadsheet program.

Table 3: When  $\lambda = -0.06$ , Expected Gain in Followers from Change in Confidence and Accuracy

Base Level of Followers	Gain from Confidence ( $C_i$ )		Gain from Accuracy ( $C_i$ )	
	50% Confident	100% Confident	50% Accurate	100% Accurate
250	8%	17%	18%	39%
1000	9%	19%	19%	43%
2500	10%	20%	21%	46%
5000	10%	21%	22%	48%

The above table shows the percent gain in followers associated with the supplier displaying 50% and 100% confidence or accuracy. These values are based on a regression where the dependent variable has underwent a Box-Cox transformation equal to the estimated fit ( $\lambda = -0.06$ ).

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
1	twitterid	score	verified	following	followers	user_created	status_count	listed_count	favorites_count	tweet_created	retweet_count	is_retweet	inconversation	account_age	tweet_flow	t_followers
2	6.137E+17	1	0	950	1232	2014-06-02	24397	9	0	2015-06-24	0	0	0	1.05942309	23028.5712	1232
3	6.137E+17	1	0	950	1232	2014-06-02	24397	9	0	2015-06-24	0	0	0	1.05942309	23028.5712	1232
4	6.137E+17	1	0	1864	2077	2012-08-11	12715	36	166	2015-06-24	0	0	0	2.86651507	4435.69969	2077
5	6.137E+17	3	0	1760	1691	2012-10-28	4922	82	965	2015-06-24	0	0	0	2.65291439	1855.31807	1691
6	6.137E+17	3	0	920	723	2014-09-29	287593	111	0	2015-06-24	0	0	0	0.73291445	392396.412	723
7	6.137E+17	1	0	4006	3759	2011-09-14	3259	27	1632	2015-06-24	0	0	0	3.77464665	863.392075	3759
8	6.137E+17	2	0	148	597	2012-02-06	170626	25	112	2015-06-24	0	0	0	3.3786928	50500.5859	597
9	6.137E+17	3	0	62	9	2015-06-23	5	0	0	2015-06-24	0	0	0	0.00354447	1410.64771	9
10	6.137E+17	3	0	199	154	2011-01-25	385	21	15	2015-06-24	0	0	0	4.41095682	87.2826499	154
11	6.137E+17	3	0	118	120	2015-05-04	4659	51	2028	2015-06-24	0	0	0	0.133961401	33370.5771	120
12	6.137E+17	3	0	5786	11277	2009-07-18	10281	103	11494	2015-06-24	0	0	0	5.93398072	1732.56377	11277
13	6.137E+17	1	0	44	136	2009-08-27	1465	1	2	2015-06-24	0	0	0	5.82340045	251.571228	136
14	6.137E+17	1	0	358	3177	2011-02-24	2814	66	3	2015-06-24	0	0	1	4.32888623	650.051734	3177
15	6.137E+17	1	0	74	264	2011-05-21	305	4	47	2015-06-24	0	0	0	4.09365058	74.5054443	264
16	6.137E+17	3	0	1772	1236	2015-02-18	86318	383	7	2015-06-24	0	0	0	0.34633312	249234.034	1236
17	6.137E+17	3	0	860	916	2013-12-30	18596	342	780	2015-06-24	0	0	0	1.48214364	12546.6922	916
18	6.137E+17	3	0	360	385	2009-02-03	3858	35	108	2015-06-24	0	0	0	6.38350873	604.369817	385
19	6.137E+17	1	0	11631	12507	2012-03-17	27178	337	6	2015-06-24	0	0	0	3.27023357	8310.72136	12507
20	6.137E+17	3	0	642	963	2008-04-05	12445	58	232	2015-06-24	0	0	0	7.21805793	1724.14798	963
21	6.137E+17	3	0	771	504	2008-10-07	1868	64	311	2015-06-24	0	0	0	6.71070287	278.361304	504
22	6.137E+17	1	0	818	1608	2011-03-02	1296	22	0	2015-06-24	0	0	0	4.31202924	300.554548	1608
23	6.137E+17	1	0	7381	18432	2009-02-10	72384	175	11975	2015-06-24	0	0	1	6.3650931	11372.0253	18432
24	6.137E+17	3	0	211	93	2012-08-02	309	0	2	2015-06-24	0	0	0	2.89165095	37.6947294	93

Figure 5: The output of the analyzer as shown by Microsoft Excel. The output includes the tweet identifier, information on the writer of the tweet and the score from the regular expression CSV file.