The 2012 Wisconsin Gubernatorial Recall Twitter Corpus

Christopher Mascaro, Alan Black, Michael Gallagher, Sean Goggins
Drexel University

Abstract
In the following document we detail what we identify as the Wisconsin Gubernatorial Recall Twitter Corpus. This dataset represents two and half months of collection surrounding and incorporating the June 2012 Wisconsin Gubernatorial Election collected using the Group Informatics Lab Infrastructure TwitterZombie at Drexel University (Black, Mascaro, Gallagher, & Goggins, 2012). This document is a précis for the larger dataset. We intend this document to be a way share with other researchers the type of syntactical feature use we identified in a statewide political race that many deemed to have national implications in a General Election year. Through this sharing we hope to build on existing knowledge about political discourse on Twitter and we welcome any researchers who are interested in collaborating in this space.

Context

Governor Scott Walker was elected to the Wisconsin Governor's office in November 2010 along with a large number of national Tea Party candidates who stood for leaner government and more controlled spending. In early 2011, Scott Walker backed the 2011 Wisconsin Act 10, that aimed to fix significant gaps in the Wisconsin operating budget. The bill attempted to eliminate collective bargaining rights for Wisconsin government employees and proposed that government workers contribute more to their state pension plans.

Immediately following the introduction of the bill, widespread protests occurred in the state capitol of Madison estimated to number 100,000 individuals at their peak. This lead to fights between protesters and officials that eventually lead to significant restrictions being placed on who could enter the capitol building and also lead to 14 Democratic Senators exiling themselves in Illinois, an adjacent state, to prevent a quorum being present for the vote. Under Wisconsin law, they could be compelled to vote if they remained physically in the state. The ensuing protests lead to a wave of protests throughout the United States as collective bargaining rights were seen threatened by many. The bill was passed and signed by Governor Walker on March 11, 2011.

The passing of the bill seemed counter to Scott Walker's original campaign platform, which was more moderate, and this lead to a movement to recall him. In March 2012, Wisconsin officials announced that opponents of Governor Walker had met the requirements for a recall election. The Democratic Primary between Milwaukee Mayor Tom Barrett, Kathleen Falk, Secretary of State Doug La Follette and State Senator Kathleen Vinehout took place on May 8, 2012 with Mayor Barrett winning with 57% of the vote.

The month long campaign that preceded the General Recall Election on June 5, 2012 was marked by attacks from the Mayor Barrett that Scott Walker had been involved in improper campaign financing. Governor Walker attacked Mayor Barrett with allegations that his police department underreported violent crimes to make his policies appear to be more effective. These attacks were made most salient in two debates that occurred on May 25 and May 31. In the second debate on May 31st, Mayor Barrett said when talking about Governor Walker: "I have a police department that arrests felons. He has a practice of hiring them." This line brought more national attention to the election and appeared to have given Mayor Barrett an uptick in support. This support was not sustainable as Governor Walker defeated Mayor Barrett a few days later by a margin of 53%-46%.

Dataset

Collecting big data requires diligence and a careful selection of the terms that are collected to ensure a valid, yet comprehensive dataset that allows the researcher to answer the research questions they have posed. The collection strategy for this dataset is described in detail so that researchers can assess what the data represents in the context of Twitter and online political discourse more broadly. We do not intend

this dataset to be representative of everything that was discussed on Twitter related to the election, but we believe that our extensive search terms coupled with our documented collection (Black et al., 2012) and analytical infrastructure (Mascaro & Goggins, 2012) that it represents a significant portion.

The complete dataset consists of Tweets collected using two collection mechanisms. The first dataset is a collection of 16 Wisconsin Gubernatorial recall election specific queries collected using the Group Informatics Lab's TwitterZombie collection infrastructure (Black et al., 2012). In the TwitterZombie infrastructure, each query is known as a job and the terms job and query are used interchangeably throughout our lab's research. The second dataset are the candidate's Twitter timelines representing the collection of tweets sent out by a candidate's Twitter user name (also called a "handle"). These tweets were collected using the twitteR package (Gentry, 2012) in the R statistical computing environment (Team, 2012).

There are three distinct events that are reflected in the data. First, the initial discourse surrounding the recall election and the Democratic primary that occurred on May 8. Second, the overall discourse pertaining to the recall election and the general election between Tom Barrett and Scott Walker that took place on June 5. Since the primary election and general recall election all pertained to one event and the discourse was situated in a larger context that called for the recall of Walker, the hashtags and other discourse markers for the two elections are the same.

The third event reflected in the dataset is related to the two debates that occurred between Tom Barrett and Scott Walker. The discourse on Twitter for these two debated was marked through the utilization of the #widebate hashtag. An initial examination of this data has already been published, but more work is currently being done (Mascaro, Black, & Goggins, 2012). Similar to our analysis of backchannel debate discourse (Mascaro & Goggins, 2012) our analysis of this discourse identifies distinctly different behavior of users between the two debates.

TwitterZombie Data

The queries that were collected using TwitterZombie were a combination of hashtags, keyword queries and candidate Twitter handles that were mentioned by other users. The first set of queries that were started on April 2, 2012, were based on an original set of hashtags and keywords identified as being associated with the recall. Queries were identified by searching Twitter when the recall was first announced and identifying the most frequently occurring hashtags associated with "Wisconsin and Recall." This search was frequently conducted to identify any new queries.

In late April, analysis of the collected data highlighted another set of hashtags and queries that were surfacing in collection. These queries were mostly associated with the Democratic primary to determine who was to face Scott Walker in the recall election. In addition to those queries, #widebate, a hashtag associated with the two debates was also collected starting in late May right before the first of two debates. This phased set of snowball sampling (Figure 1) based on existing data allowed for a more complete set of data as new discourse markers were instantiated and adopted through the course of the election.

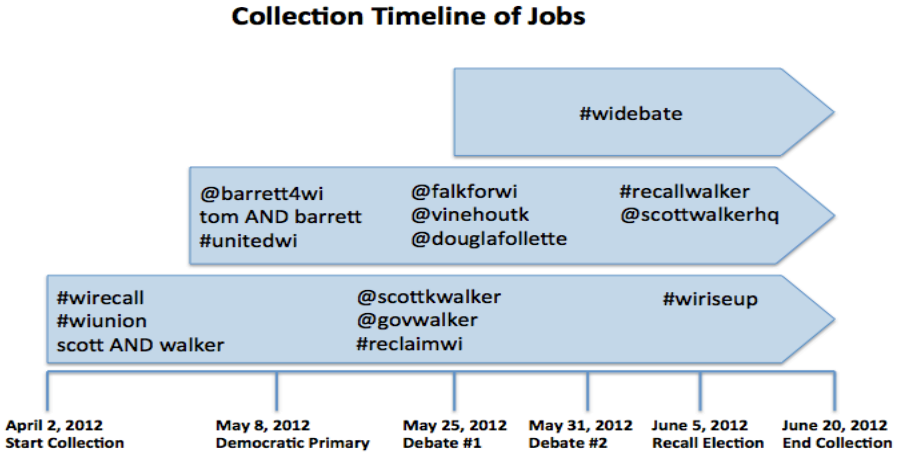

Figure 1: Syntactical feature collection timeframes

The only two keyword queries that were tasked were the general election candidate names ("scott AND walker" & "tom AND barrett"). We did not choose to task any queries such as "Wisconsin AND recall" as an initial analysis of the data that was being pulled in Twitter revealed a significant number of spam messages and those messages that did not contain spam contained another term that would be selected on using another query such as the hashtag #wirecall. Although it is possible that some tweets were missed as a result of not tasking those terms, it is believed that this set is limited and that the dataset that was collected is still representative of the discourse on Twitter related to the election.

Table 1 identifies each of the jobs along with the total number of tweets that were associated with each query and when the first tweet associated with each job was collected. In total, there were 1,084,560 tweets collected. Since the TwitterZombie infrastructure collects a tweet if it meets a certain criteria, it is possible that a single tweet could be collected by multiple "jobs." For example, a tweet that contained "#wirecall," "#wiunion" and "@govwalker" would be counted as a tweet in all three jobs and associated with each job in the database. This is done as a feature of the infrastructure to overcome the "bucket of tweets" problem that makes disaggregation difficult. Therefore, the total number of tweets when pulled down by each job is likely to include many duplicates.

To create a dataset where each tweet was represented only once, an R script was written to remove duplicate tweets using the unique tweet id number that is collected with each tweet by TwitterZombie. The R script created two datasets, one with each unique tweet (n=860,143) and one with the tweets that were deemed to be duplicates in the dataset (n=224,417). To verify the function of the R script that identified the duplicate tweets, another R script was written to generate a random sample of 1% of the tweets for further analysis. Analysis of the random sample of duplicates verified that the tweets identified as duplicates fit into two or more job queries and therefore would have been associated with multiple jobs.

Job/Query	Tweets	Overall Percentage
#wirecall	244,338	22.5288%
#wiunion	214,401	19.7685%
scott AND walker	324,889	29.9558%
@scottkwalker	43,650	4.0247%
@govwalker	81,889	7.5504%
#reclaimwi	8,110	0.7478%
#wiriseup	481	0.0443%
@barrett4wi	35,573	3.2799%
tom AND barrett	52,708	4.8599%
#unitedwi	842	0.0776%
@falkforwi	1,080	0.0996%
@vinehoutk	172	0.0159%
@douglafollette	47	0.0043%
#recallwalker	48,260	4.4497%
@scottwalkerhq	2,010	0.1853%
#widebate	26,110	2.4074%
Total	1,084,560	

Table 1: Tweet Count by Job

All of the queries were stopped at 1500 GMT on June 20, fifteen days after the election. This date was chosen as the number of new tweets per day had fallen to under 1,000 (Figure 2) and over 50% of these new tweets were associated with one the #wiunion query (Figure 3). The hashtag #wiunion deals with larger union issues in Wisconsin and was used during the Wisconsin recall event in combination with other

hashtags as a way to discuss the primary issue which motivated the recall election, the attempted elimination of state employee unions. Examination of the new tweets that were being collected on a daily basis also highlighted that the discourse had shifted from the election on Jun 5 to placing the Wisconsin Recall election in the context of the General Election in November.

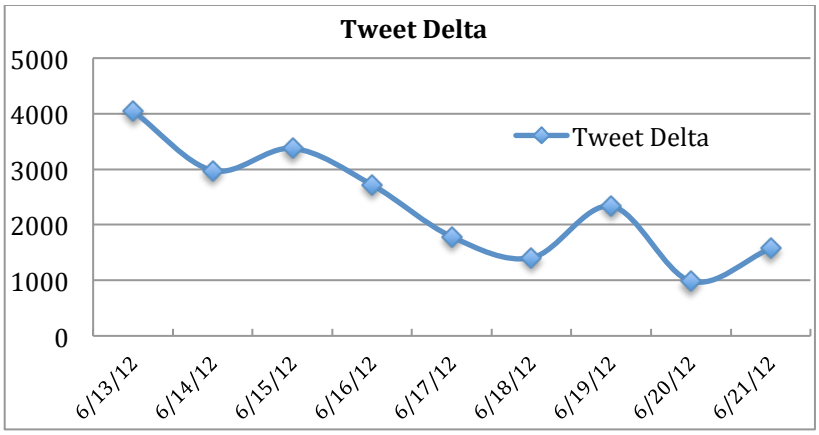

Figure 2: New Tweets by Day

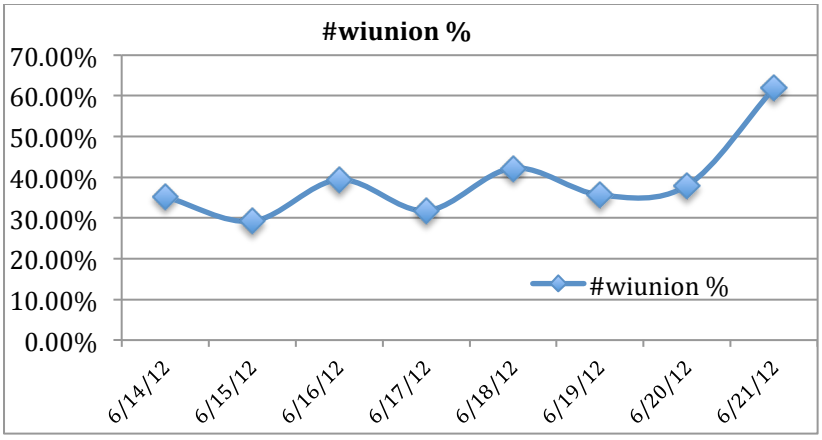

Figure 3: Percentage of #wiunion new collection

Twitter Timeline Data

The second set of data that were collected for analysis consists of timeline data collected using the twitteR package in R. Since TwitterZombie uses the search API to collect tweets, it does not collect tweets made by specific users, it only collects tweets mentioning the users that are tasked in the system. Therefore, twitteR was used to collect the timelines (all Tweets made by a particular Twitter user with a public profile) of Scott Walker ("@scottwalkerhq" (campaign), "@scottkwalker" (personal), "@govwalker" (official)) and Tom Barrett ("@barrett4wi" (campaign)). This allows for the comprehensive picture of tweets related to the candidates, which includes what they are saying and what people are saying about them.

The two sets of data that were collected represent different, but complementary types of activity. The activity from TwitterZombie illustrates discourse that originates from citizens and news organizations pertaining to the recall election. The timeline data reflects the public face of Scott Walker and Tom Barrett. When placed in the context of the other discourse collected by TwitterZombie, the candidate timelines illustrate agenda setting and response behavior to the public. Combining the citizen discourse with the discourse of the candidates and their profile information provides a multifaceted set of data related to the activity in relation to the candidate's use of Twitter.

Known Data Gaps

It is important to note that a technical problem with the collection system limited the collection of some of the tweets during a 90-minute period in the late morning on June 6. This technical outage was the result of a combination of network connectivity issues at Drexel, software performance issues and limitations inherent in the Twitter Search API. These types of collection issues are an issue with all Twitter based collection. The advanced monitoring and logging functions in TwitterZombie enable us to report these limitations with a specificity not reported in other publications analyzing Twitter data. This advanced infrastructure, combined with trending data on the rate of tweets before the downtime and the fact that TwitterZombie collects the 1,500 most recent tweets, enables me to report with confidence that less than 5,000 tweets were missed during this period.

This represents less than one-half of one percentage point of the overall unique tweets that were collected and analyzed as part of this pilot study. The fact that so few tweets are estimated to have been lost and that the election outcome was known the night before, leads to an empirically verifiable assessment that the tweets missed do not undermine the validity of the work. The specific knowledge of missed tweets available from the infrastructure, in fact, enables me to address the research questions with an uncommon degree of clarity about the nature of the data set. . The only analysis impacted by this outage is related to time series analysis, which contains a gap as a result of the outage. It is for this reason that I do not explicitly use a granular time series analysis as a method to examine my data and instead partition the complete time period into larger time periods for examination. This should limit the effect of this gap in collection.

Syntactical Feature Overview

In an effort to narrow the conceptualization of the various syntactical features of Twitter we employ a narrow conceptualization of the most common syntactical features and report on their prevalence in our dataset. We recognize that there are many other ways for these types of syntactical features to be utilized by users, but we attempt to use the most common ones that are supported by most technology. Table 2 specifies these conceptualizations clearly with the syntax we use along with a brief explanation of the overarching purpose of the feature.

Syntactical Feature	Common Syntax	Purpose
@-Reply	@[username] at first position of tweet text	To directly address another individual in a public manner
Mention	@[username] at any point in tweet text	To highlight a tweet to another individual or to talk about someone. Mentioning them will inform them of the tweet
Retweet	RT @[username] “tweet text”	To further disseminate another individual’s tweet.
Links	http://t.co/[8 characters]	To include external information in a tweet. Note: Twitter uses a URL shortener
Hashtags	#[alphanumeric text]	To tag a message with a conversational marker or to add a tweet to an existing stream of discourse independent of a follower/followee network

Table 2: Syntactical Features Enumerated

Table 3 illustrates the frequency in of certain types of syntactical features and also provides general descriptive statistics of the dataset. We report the percentages based on the narrow

conceptualization of these syntactical features as identified in Table 2. We further report the number of unique participants, number of unique tweets (retweets would be counted as duplicates), along with the number of single participants and number of tweets in English. We also report the mean tweet character length to demonstrate that the average tweet uses a significant amount of the available character space for the tweet.

Link %	Hash %	Mention %	@-Reply %	Retweet %
46.61%	66.11%	71.26%	8.01%	52.25%
Unique Tweeters	**Unique Tweets**	**Singleton %**	**English %**	**Mean Tweet Character Length**
143,504	546,088	55.70%	98.40%	118.3

Table 3: Complete dataset

Governor Walker and Mayor Barrett's Public Twitter Personas

The analysis up until this point has examined how individuals in Twitter talked about the candidates and the election. One of the benefits of Twitter's API is to be able to not only collect what individuals say using specific syntactical features, but also collect the timelines of specific users. The timelines of users allow for the examination of the public face of Twitter for that individual. In the case of political candidates, this public face is a significant part of a campaign.

There are 4 accounts that I collected the timelines for: @scottwalkerhq (campaign), @scottkwalker (personal), @govwalker (official) and @barrett4wi (campaign). Mayor Barrett did not have an account affiliated with his official duties as mayor of Milwaukee at the time of collection in late June 2012. Table 3 denotes the total number of tweets from each of the accounts from March 5, 2012 – June 20, 2012. The number in parentheses is the number of tweets that were tweeted by each of the accounts after the recall election was officially announced in late March. The start date of the overall collection of these handles was determined as this was the first public tweet collected from Governor Walker's campaign account and the end date was determined as it was the end of my data collection that was determined in the earlier section. The inclusion of the second number in the tweets columns is to get a better sense of how the activity evolved over time.

User	Tweets	% in dataset
barrett4wi	269 (248)	89%
govwalker	546 (378)	2%
scottkwalker	322 (283)	7%
scottwalkerhq	243 (188)	59%

Table 3: Candidate Tweets by Account

The third column of table 3 shows the number of tweets of each of the candidate accounts that would have been visible in the dataset if one did not also collect the timelines. This percentage reflects the candidate usage of the syntactic features that were being collected and identified in table 1. We see that the two highest percentages are for those accounts that are directly related to the campaign for each of the candidates. In the case of Governor Walker, his usage of his official account for the recall election was extremely limited. One of the sub-findings of our overall analysis is related to the collection posture. It is important to collect both the timelines and @-mentions of individuals of interest as their message may differ from the public discourse about them. Examining the two sides allows for a more comprehensive analytical approach and provides ground truth of Twitter activity from individuals of interest.

To examine how the two candidates managed their public personas, it is important to do an in-depth analysis of a set of their tweets. Since Mayor Barrett and Governor Walker both maintained a campaign account and the topic of interest is the election, an analysis of the tweets from their campaign accounts is an appropriate comparative analysis for our purposes. This selection criterion is further supported by the low percentage of tweets directly related to the recall election identified for @govwalker and @scottkwalker's accounts in table 3. Table 4 represents the syntactical feature breakdown of the two campaign accounts of Mayor Barrett and Governor Walker.

Handle	Link %	Mention %	Hashtag %	At-reply %[1]	Retweet %
@barrett4wi	69.35%	48.79%	64.52%	10.08%	0.81%
@scottwalkerhq	67.76%	49.18%	35.52%	8.20%	0.00%

Table 4: Syntactic Feature Usage of Candidate Account

Conclusion

This documents gives a high level overview of the syntactical features utilized to engage on Twitter surrounding a statewide Gubernatorial race. The Group Informatics lab is currently developing articles that examine many of the specific phenomena observed in the data including retweet and URL utilization, conversational discourse and hashtag evolution. We welcome other researchers from any field who are interested in collaborating in the analysis of the dataset. In addition to this dataset, we also have been collecting data for the 2012 Congressional races and the eleven other Gubernatorial races along with the Presidential election that we encourage collaborations on.

References

Black, A., Mascaro, C., Gallagher, M., & Goggins, S. (2012). *TwitterZombie: Architecture for Capturing, Socially Transforming and Analyzing the Twittersphere*. Proceedings from ACM Group, Sanibel Island, FL.

Gentry, J. (2012). *twitteR package for R.*

Mascaro, C., Black, A., & Goggins, S. (2012). *Tweet Recall: Examining Real-Time Civic Discourse on Twitter*. Proceedings from ACM GROUP, Sanibel Island, FL.

Mascaro, C., & Goggins, S. (2012). *Twitter as Virtual Town Square: Citizen Engagement During a Nationally Televised Republican Primary Debate*. Proceedings from American Political Science Association Annual Meeting, New Orleans, LA.

R Development Core Team. (2012). *R Foundation for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

[1] About 80% of Tom Barrett's at-reply's began with a period. This is a specific Twitter syntax that ensures that an at-reply is seen publically. None of Scott Walker's at-reply's contained this syntactical structure. The usage of the syntax has not been widely studied to this point and since the point of using the period before @ is to still directly address someone else, we do not identify this as a different form of syntactical feature.