The Method to the Madness: The 2012 United States Presidential Election Twitter Corpus

Christopher Mascaro
Drexel University
Philadelphia, PA
cmascaro@gmail.com

Denise Agosto
Drexel University
Philadelphia, PA
dea22@drexel.edu

Sean P. Goggins
University of Missouri
Columbia, MO
GogginsS@missouri.edu

ABSTRACT
Social media provides a rich environment for understanding social connections, interactions and information sharing across many aspects of society. The relative ease of access to social media data through provision of APIs by the companies has led to a significant number of studies that attempt to understand how social media fits into society and how the public uses it for discourse and information sharing. One of the existing gaps in these studies is the lack of extensive description of the data collection and processing methods. These gaps exist as a result of word limits in existing publication venues and a lack of appropriate publication venues to share this type of fundamental research. The following paper provides extensive detail as to how a 52 million corpus of Twitter data on the 2012 Presidential Election in the United States was collected, parsed and analyzed. This level of detail is imperative in studies of social media as small choices in what data to collect can have material effect on the findings. In addition to the description of the methods, the following paper provides a contribution to knowledge in providing basic characteristics of one of the largest research datasets of social media activity compiled to study political discourse.

CCS Concepts
Information systems → Social networking sites

Keywords
Twitter, social media, political, information technology, election

1. INTRODUCTION
Collection of electronic trace data should be theoretically informed and take into account the topical and technological context [8-10; 21]. Social media is a rich source of data, but in order to derive scientifically sound findings, it is imperative to describe data collection processes and the provenance of the data.

Existing word limits in many publication venues make the extensive description of specific methods used for the collection of data for each study nearly impossible. The following paper helps guide the reader through the complete research process from dataset construction through the findings and discussion of a dataset that forms the basis of numerous studies that are currently under review or accepted for presentation in other venues [20].

In studies that use traditional methodological approaches such detail may be superfluous, but with the evolving APIs, tools and methodological approaches used in social media research, transparency is required to help others understand not just the methods used to examine the data, but also how the dataset was constructed as each decision is material to the overall findings [17].

Previous research that has focused on methodologies to collect Twitter data has attempted to understand how collection decisions affect results, but these studies lack contextual specificity and instead attempt to characterize different technologies that can be used to collect the data [1; 5; 11; 22; 27]. The primary finding of these studies is that the choice of API or collection tool along with the specific type queries used to collect data have material affect on findings. In the following paper, the choice of API is discussed, but the focus is on how the dataset is constructed through theoretical and contextual query selection along with specific description of the conceptualizations of medium specific features.

The following paper presents a comprehensive overview of the methods and decisions made to construct a dataset that captures the activity within one technological medium, Twitter, during the 2012 Presidential election. The constructed dataset does not represent every tweet that occurred during the specified timeframe that was related to the election, but the explicit description of the selection and collection criteria allows the reader to understand the manner in which the dataset was constructed and understand how the findings of subsequent papers were derived [20].

The combination of the time period that collection occurred coupled with the constructed queries leads to the construction of a dataset focused on political discourse. Political discourse as conceptualized by Van Dijk [29] and others [6; 7] represents discourse about politicians, among politicians and about issues that are of concern to politicians. The dataset represents one of the largest datasets of social media data used to understand how political discourse is constructed and evolves over the course of a campaign.

The current published literature that discusses the 2012 Presidential Election on Twitter has focused on sentiment analysis of tweets [30], the relationship between mentions and candidate activity in the primary [13], content analysis of candidate tweets [15], retweet analysis of minority party candidates [4] and how Twitter was used for agenda setting [19]. In addition to studies of Twitter similar studies of other technologies such as Facebook were also conducted [2].

None of the published studies include an explicit description of data collection methods needed to understand the implications of findings nor do they take the comprehensive approach of collecting data using a diverse set of queries. This limits the comprehensive understanding of the findings and context needed to understand the activity that is occurring within the technological medium.

Social Media & Society'16, July 11–13, 2016, London, England.

The dataset that is described in the following paper provides the foundation of a set of papers that identify several distinct contributions to knowledge that address the following research questions using a combination of network analysis, co-occurrence analysis and temporal frequency analysis to examine the 53 million Twitter message (tweet) corpus collected during the 2012 Presidential Election (August 20, 2012 – November 13, 2012)[1].

1. How does the political discourse related to the 2012 Presidential election manifest itself in a technologically mediated environment?
    a. How does this discourse in Twitter differ during acute events such as debates compared to long-term discourse that occurs throughout the election?
    b. How does the use of politically oriented hashtags in Twitter facilitate this discourse?
2. How does the political discourse that occurs in Twitter using the syntactical features of the at-mention and at-reply surrounding the 2012 Presidential Election identify an emerging participatory public engaged in political discourse?
3. To what extent do the URL and retweet syntactical features in Twitter facilitate information exchange surrounding the 2012 Presidential Election?

In addition to the description of the methods used to construct and collect the dataset, the following paper provides a summary of the key characteristics of the activity that occurred within on Twitter during a specified time period. The statistics that detail the number of times certain syntactical features of Twitter were used constitute significant contributions to knowledge in their own right and also provide the ability to compare across datasets the distribution of the percentage of tweets that contained hashtags, URLs, at-mentions and at-replys identified in a dataset of political discourse [25].

2. DATA COLLECTION

The data collection for this study was done using a modified version of the TwitterZombie [1] infrastructure. The software that conducted the management of the queries and collection was the same as the original TwitterZombie infrastructure that was previously validated [1]. Given the nature of the collection requirements for what was anticipated to be a high volume event (in this case the 2012 Presidential Election), the technical infrastructure of the original TwitterZombie system was redesigned.

2.1 Original TwitterZombie design

The original TwitterZombie infrastructure utilized one collection head that stored data in one database. This technical architecture was deemed to be insufficient for high volume collection for two reasons. First, Twitter limited the amount of tweets that can be collected at one time for each query using the SEARCH API to the 1,500 most recent tweets. This constraint limits the overall number of tweets that can be collected by each job during each collection run. In turn, this requires jobs with high volume to be run frequently in order to collect the data. Second, each query can take anywhere from 2-45 seconds to run in the TwitterZombie system based on the volume of data that is being collected and each query must finish before the next one begins[2].

One of the design decisions for the software architecture used in the original TwitterZombie was to run queries sequentially. Once a query finished, the next would begin. This design decision allowed for TwitterZombie to recover from any system or network interruptions or technical issues that may be present in the Twitter network. If one query failed, the next would execute successfully because there was no dependency. Although this design decision increased the resiliency of TwitterZombie, it also limited some of its applicability to high volume events. In combination, these two limitations were deemed to be prohibitive for collecting a comprehensive dataset that used multiple queries in a narrow timeframe during a high volume event.

During the initial testing and development of the TwitterZombie system [1] it was identified that high volume events would overburden TwitterZombie as it was possible that a job would take too long to complete. This would delay the collection of another job that was next in sequence. For example, if three jobs were supposed to collect every minute on one head and the first two jobs ran over one minute, the third would start collecting and delay the first two jobs from running on schedule. This would create a cascade effect that would lead to a collection gridlock that may never be overcome leading to significant gaps in collection. Therefore, a new technical infrastructure was developed to account for high volume events.

2.2 Redesigned Technical infrastructure of TwitterZombie-n

As a result of this need for high volume collection, the TwitterZombie technical infrastructure was rebuilt using the Amazon Web Services (AWS) cloud infrastructure. The cloud infrastructure provided collection elasticity that allowed for scalability based on collection requirements such as the occurrence of an acute event such as a debate. The new system, "TwitterZombie-n", was an adaptation of the original Twitterzombie infrastructure for the cloud environment. In the new infrastructure, multiple collection heads were used to collect data. This allowed for the maximum possible collection based on the available resources given financial and technical restraints and for a higher volume of data to be collected. Although there are multiple collection heads in the new design, the infrastructure was built using one central "job" table[3]. The centralization of the job table limited the possibility of system failure at one of the collection points and provided one systematic point of control of collection priorities and queries.

The distribution of multiple queries across multiple collection heads limited the possibility of collection leading to overburdening heads as jobs were evenly distributed among heads. In the early part of the election cycle, four heads were utilized, but this was expanded to five during October and November to account for higher traffic volume from the conventions, debates and election day as evidenced in the Figure 1. The system was still limited by the limit of 1,500 tweets per minute for each job, but the distribution of queries amongst the collection heads allowed the system to run as efficiently as possible given its tasking.

[1] These papers are currently under review at a number of other venues.

[2] This limitation is not specific to the TwitterZombie system and merely reflects the delay that would exist for collection between any collection system and the Twitter API.

[3] In Twitterzombie, a job and a query are synonymous and used interchangeably.

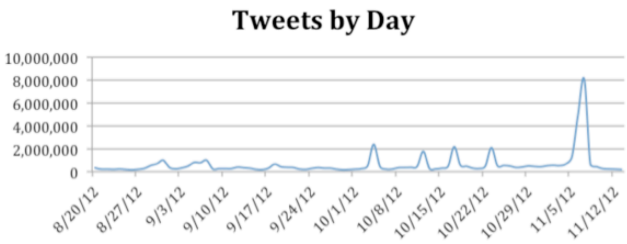

Figure 1: Tweets by Day

The TwitterZombie-N collection system was run at an Amazon Web Services colocation facility in Northern Virginia for the duration of the collection cycle (August 20, 2012 – November 13, 2012). In an effort to eliminate any possibility of collection failure during high volume events, another duplicate system was run during the last 2 weeks of the election cycle (October 28 – November 13) in another AWS facility in Oregon. The geographic separation assured availability of resources to limit any collection faults that were the result of a loss of network activity or other unanticipated technical failure[4]. This redundancy ensured that a dataset that covered the whole time period without any gaps would be collected

As a result of the high volume collection of many of the queries during the last two weeks of the election, it was decided to combine the two datasets (Northern Virginia and Oregon) into one dataset. This combination was done using the "tweet_id" string that is included as part of each tweet to ensure no duplicate tweets existed in the combined dataset. As a result of the addition of the Oregon dataset to the Northern Virginia dataset an additional 2 million tweets were collected[5].

These additional tweets are the result of high frequency Twitter activity during the conventions, debates and on Election Day that the two systems (Northern Virginia and Oregon) were able to collect based on a slight offset of clocks on the two systems. Black et al. [1] identified the possibility for differences in collection based on the time and this phenomenon was identified in the TwitterZombie collection infrastructure. Although the job tables were offset by a matter of seconds, they collected a number of different tweets during high volume events. This redundancy and elasticity helped to ensure a more representative collection sample, even though there are likely data gaps as a result of limits placed on public API access by Twitter[6].

3. DATASET CONSTRUCTION

The TwitterZombie architecture was designed to limit the need to disaggregate collected data. In the system, each tweet that meets the selection criteria for a query is collected as a distinct entry in the database. Therefore, the tweet: "*@BarackObama Good work in the debate Mr. President beat @MittRomney #debates #election2012"* would be collected five times by the TwitterZombie jobs that were currently collecting data for the election (@BarackObama, @MittRomney, #debates, #election2012 and "barack AND obama").

This design allows for easy access to tweets that meet certain criteria, as it is possible to query the database by "job number" as opposed to querying the database for a string of characters. On one hand, the discrete job selection is powerful as one is able to get data quickly when examining tweets for a specific query. On the other hand, the creation of a dataset that represents the discourse about a larger topic that traverses multiple jobs during a specified time period becomes costly in terms of processing time. In turn, this approach requires a clear set of research questions and description of why specific jobs were included in the larger dataset. The following section describes the construction of the dataset for this study in terms of three categories of selection criteria: event related queries, candidate related queries and queries related to campaign promoted activity in Twitter.

The dataset used for this study was the result of combining 68 theoretically informed queries that are described in the next section. The complete dataset after the combination of Northern Virginia and Oregon datasets and the 68 queries amounted to 62,806,682 tweets. After using the SQL *distinct* command on the unique tweet identifier (tweet_id) of each message to eliminate duplicates that would occur as a result of multiple syntactical features being present in a tweet, the final dataset amounted to 52,487,179 tweets. This is a reduction of 10,319,503 tweets, approximately 16% of the original dataset.

The 16% overlap highlights the fact that search queries were highly correlated as one tweet was collected by multiple queries. This is likely a result of the theoretically informed query selection [8] that was done to ensure the widest amount of political discourse related to the 2012 Presidential Election. It is intuitive that tweets about President Obama during the election would also likely mention a debate or his opponent Mitt Romney and would be collected by multiple jobs resulting in overlap. The next section identifies the selection of jobs and inclusion of many queries that were used to collect political discourse during the time period of interest to afford a level of transparency in the construction of the dataset [17].

3.1 Job Selection

In a big social data environment a representative sample of data to address stated research questions is desired. Unfortunately, this type of sample is difficult to collect as it is unclear as to whether the technology provider is limiting access to data or whether individuals are using appropriate terms for comprehensive collection. This is a problem that is not only a factor with this study, but also with all studies that rely on API-based collection from technology services that are not under the control or management of the researcher or without other data sharing relationships established between the researcher and the company.

To date, there has been limited analysis of the specifics of the different Twitter APIs and this is definitely an area for future research, but outside the scope of this work [1; 5; 11; 22-23; 27]. In an effort to collect as much political discourse related to the election during the time period, a total of 68 queries were collected on in the Twitterzombie-n infrastructure. These queries were a combination of Twitter handles, hashtags, and keywords

[4] The AWS location in Northern Virginia where the original system was being run had multiple documented occasions in the previous year where service was lost for a multiple hour time period. This led to the loss of availability of popular websites such as Netflix. The addition of the AWS site in Oregon was intended to guard against this possibility and in turn led to greater collection.

[5] An analysis of the tweet_id strings was conducted to identify any duplicates after the datasets were combined and there were no duplicates identified.

related to the candidates and temporal events of the election such as the conventions, debates and Election Day.

The initial set of queries was established by identifying all of the handles associated with the candidates and the campaigns, including the wives of the candidates since they were also participating in campaign events. In addition to these handles, the first and last name of the candidates were added to the collection in an effort to capture anytime an individual references one of the candidates without using their Twitter handle. Additionally, specific hashtags that were used in the candidate's Twitter feeds were collected as these represented areas of discourse.

Throughout the data collection period, new hashtags were identified as associated with electoral events and promoted by candidates. These were not included in the initial set of queries, but were added to the collection infrastructure to attempt to collect a comprehensive dataset. All of the queries that were added during the election were kept on collection through the end of the collection period (November 13, 2012).

3.2 Candidate Dataset Creation

The queries related to the major Presidential and Vice Presidential candidates formed the foundation of the dataset (Table 1). Each of the candidate's official Twitter handles was collected as part of the collection process. In the case of Paul Ryan, his Congressional handles were also collected since he was concurrently running a Congressional campaign and it was likely that the discourse would bleed over between the two campaigns[7]. In addition to each handle, the hashtag for each of the candidate's full names, first names and last names was tasked. Hashtags for Paul Ryan were not collected due to the common nature of his name and the identification of completely unrelated discourse related to his name in an early sample of the data.

In addition to the handles for the candidates, a keyword combination of the candidate's full names was tasked to ensure collection of data for when the handle or hashtag was not used with the candidate's name. The keyword query followed the format of [candidate first name]%20[candidate last name][8]. In addition to the full names, the query "president %20 obama" was used to collect tweets that referred to President Obama by formal title instead of his name.

Twitter handles for the candidate's wives along with the official White House account @WhiteHouse were also collected. Finally, the hashtags for the Presidential and Vice-Presidential candidates and "2012" (e.g. #obamabiden2012) were also tasked to collect data from those who may have used the hashtag as an affiliation. Table 3 illustrates the queries for the candidates and hashtags that represented campaign specific discourse.

Barack Obama	Joe Biden	Mitt Romney	Paul Ryan
@BarackObama	@Joe_Biden	@MittRomney	@PaulRyanVP
@Obama2012	@JoeBiden	#MittRomney	@RepPaulRyan
#BarackObama	@VP	#Romney	@PaulRyanPress
#Obama	#JoeBiden	mitt and romney	@PaulRyan
barack and obama	#Biden		paul AND ryan
President AND obama	joe and biden		

Table 1: Candidate Queries

3.3 Event Dataset Creation

In addition to the candidates there were a number of emergent events where hashtags evolved and were used to mark specific discourse. The #election2012 hashtag was used throughout the time period and was also used before the studied time period. In total, #election2012 was used in 1,632,995 tweets in the dataset. The #electionday hashtag emerged on November 6 and was used in 221,610 tweets along with #vote (562,937 tweets), #govote (95,841 tweets) and the keyword syntax of "voted for" (1,476,786 tweets), where individuals were identifying who they voted for publicly on Twitter.

Events such as the conventions and the debates often did not have a predetermined hashtag or set of hashtags until right before the event and were difficult to collect until the day before the event. In the case of the debates, Twitter identified the official hashtag of #debates the day before the first debate even though some press releases for the first debate indicated that the hashtag would be #dudebate or #denverdebate as opposed to #debates. The use of one specific hashtag for all of the debates allowed individuals to be aware of the hashtag to identify in and participate in discourse for an extended amount of time. The use of the same hashtag for multiple discrete, yet related events (each debate) made the possibility of differentiation of discourse between debates difficult. Figure 2 illustrates that the usage of the #debates hashtag was highly concentrated the days immediately surrounding the debate making time a valid unit of analysis for examining debate specific discourse.

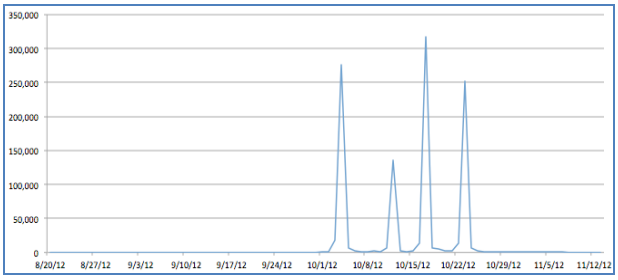

Figure 2: #debates Usage by Day

In the case of the first debate, the press release for the debate from the University of Denver identified the hashtag as being #dudebate two weeks before the debate, but this hashtag was only used 850 times. Additionally there were other hashtags identified by members of the local press such as #debatedenver and #denverdebate, but these were used in a limited capacity compared to the official #debates hashtag promoted by Twitter.

[7] Even though Paul Ryan was actively running for reelection to Congress, his Congressional Twitter account garnered limited mentions illustrating that the primary focus of activity for him during the election cycle was his Vice Presidential handle (@PaulRyanVP).

[8] "%20" represents the ASCII representation of the Boolean term AND.

Although these debate specific hashtags were not extensively used, they are included in the dataset for comprehensiveness. This behavior differs from previous analysis of primary debates where there were specific hashtags used for each debate in addition to some cable network specific hashtags such as #cnndebate or #answer and #dodge [1]. Similarly, other hashtags promoted by the local event organizers emerged during the other two Presidential debates and the Vice-Presidential debate and are included in the dataset.

Table 2 includes all of the hashtags and the respective debate that the queries were collected in relation to. Even though the hashtags were related to the specific event, collection of discourse that included these queries persisted the end of the election. This allows for the temporal analysis of the presence of the hashtag or handle. In most cases the specific debate hashtags were added to the collection system the day of or before the specific debate. It is important to note that in Figure 1, there is limited presence of #debates before the day of the debate, 10/3/2012. Although there was limited activity, collection began the week before the debate in anticipation of the hashtag possibly being of interest. This illustrates a limited adoption until the actual day of the debate.

The hashtags that were common for all of the debates were #debate, #debates, #presidentialdebate, #cnndebate, #debate2012, #presdebate and as a result of their commonality across the four debates they are not identified as being associated with a specific debate, in Table 4. The hashtags and Twitter handles that were promoted by the respective parties for the conventions are included in Table 4.

Table 2: Key Election Season Events and Related Queries

Dates	Event	Syntactical Feature Collected
August 27 - August 30, 2012	Republican Convention	#gop2012, #2012rnc, @gopconvention
September 4 - September 6, 2012	Democratic Convention	#dnc2012, #2012dnc, @demconvention
October 3, 2012	Presidential Debate #1	#dudebate, #denverdebate, #debatedenver
October 11, 2012	Vice-Presidential Debate	#vpdebate, #centrevpdebate
October 16, 2012	Presidential Debate #2	#hofdebate
October 22, 2012	Presidential Debate #3	#lynndebate
November 6, 2012	Election Day	#electionday, #govote, #ivoted, "voted AND for"

3.4 Promoted Hashtag Dataset Creation

There were a number of promoted hashtags by the campaigns. A promoted hashtag is one where an organization pays Twitter to promote a hashtag as trending on its front page. This can be done by any organization, but was used primarily by the candidates to attempt to shape the discourse on Twitter. During each morning of data collection, Twitter was examined to see if there was a promoted hashtag from a campaign. This occurred at 6 am to ensure collection began before most of the east coast had woken up. If a promoted hashtag was identified, it was added to the job table for collection.

As a result of emergent events there were a number of other hashtags that were not promoted, but were trending as a result of the large amount of discourse and events in the news that were identified in the same manner. For example, a video of Mitt Romney stating that 47 percent of Americans were reliant on government support spawned the hashtag #47percent. Table 5 identifies the promoted and emergent hashtags collected as a result with promoted hashtags have a (p) next to them[9].

Promoted Hashtags
#16trillionfail
#47percent
#romneyshambles (p)
#forwardnotback (p)
#forward (p)
#forward2012 (p)
#cantafford4more (p)

Table 3: Promoted Hashtags

4. DATASET DESCRIPTION

In total, there were 52,487,179 unique tweets that were captured using the combination of 68 queries (47 hashtags, 15 handles, 6 keyword searches). In previous research, descriptive statistics about a Twitter dataset may be biased as a result of the collection criteria. For example, a dataset collected using one hashtag and one handle would have a significant amount of hashtags and handles, but may lack any instances of an individual not using a hashtag or using an individual's last name as opposed to their Twitter handle. As a result of the selection criteria, the dataset would be biased to containing a specific type of activity.

There were 28,019,513 tweets that contained unique text. This disparity exists because a tweet that is retweeted 100 times exists 100 times in the dataset. Each retweet is its own unique tweet in the eyes of Twitter, but the actual body of the text is not unique. The comparison of unique tweet count to overall tweet count highlights that only 53.3% of the total tweets contained original text.

The most frequently occurring tweet in the dataset was a tweet from Barack Obama following his reelection that was a photo of himself and First Lady Michelle Obama hugging with the text "Four More Years." According to Twitter, this tweet was retweeted over 810,000 times in 48 hours and it appears in the dataset 394,494 times. This is three times the next highest retweeted tweet, which was retweeted 110,997 times. Until early 2014, Obama's tweet was the most popular tweet and most retweeted tweet of all time[10]. The fact that this tweet is the most frequently occurring in the dataset and was collected approximately 50% of the time that Twitter says it was retweeted provides some external validation to the collection infrastructure

[9] The trending hashtags are included for completeness of the dataset although they do not represent a promoted hashtag.

[10] As of publication, this superlative is no longer held by this tweet.

as being able to collect a significant amount of the activity even during an acute event.

It is likely that the times that this tweet was retweeted and not collected by the system occurred when the system had already collected the 1,500 maximum tweets per minute. According to Twitter, at 11:19pm ET there were 327,452 tweets per minute occurring. This rate is a record for an event in Twitter [26]. This high number of tweets per minute likely resulted in missed collection for certain queries. Since the dataset is large and most research relies on some form of a sample, the fact that the system collects on a consistent schedule indicates that the sample is representative and mirrors the activity in Twitter.

4.1 Syntactical Feature Overview

The analysis of a dataset of this size is becoming common in social media and big data research. In order to ground the methods used in this study, the conceptualizations of the syntactical features studied are in **Table 4**.

Syntactical Feature	Common Syntax	Purpose
At-Reply	@[username] at first position of tweet text	To directly address another individual in a public manner.
At-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 individuals tweet.
Links	http://[until whitespace]	To include external information in a tweet. Note: Twitter uses a URL shortener, but also accepts other URL shorteners as links too.
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 4: Syntactical Feature Conceptualization

Table 5 details the raw counts of some of the syntactical features in the dataset. There were over 48 million hashtags used, but only 1 million were unique. A similar disparity exists between the total number of at-mentions and the number of unique at-mentions. Similarly, the number of unique base URLs is less than the total number of URLs collected, since many of the base URLs were used repeatedly[11].

[11] A base URL is the domain of the URL without any of the specific directories. For example, the URL www.cnn.com/USA/news/Story, would have a unique base URL of www.cnn.com.

Unique Tweets	Total Hashtags	Unique Hashtags
28,019,513	48,083,288	1,044,858
Total At-mentions	**Unique At-mentions**	**Unique Base URLs**
55,033,314	2,617,100	118,907

Table 5: Raw Count of Syntactical Features

Examining the overall presence of certain syntactical features in the complete dataset illustrates a corpus with a diverse set of characteristics. **Table 6** provides an overview of the percentage of tweets that contained certain syntactical features. Nearly one-third of all tweets contained a URL and just over one-half contained a hashtag. Nearly three-quarters of all of the tweets contained an at-mention and this includes the nearly nine percent that were constructed as an at-reply and the 55% that were a retweet. The percentage of at-replys, at-mentions and retweets are similar to previous analyses of election data on Twitter, but the percentage of tweets with URLs and hashtags is lower in this dataset [22]. The reasons for these differences may be the addition of more keyword queries in this dataset construction and also the increased times of acute "bursty" activity where the percentage of URLs is found to be less.

URL %	Hashtag %	At-Mention %	At-reply %	Retweet %
32.59%	50.15%	72.81%	8.78%	55.10%

Table 6: Syntactical Feature Presence in the Complete Dataset

Table 7 identifies the percentage of tweets that contain an at-mention of the four candidates. President Obama (@BarackObama) was mentioned in almost one of out every six tweets whereas Governor Romney (@MittRomney) was mentioned in nearly one out of every twelve tweets. Both vice-presidential candidates were mentioned less than the Presidential candidates and these mentions were concentrated around the vice-presidential debate and election day.

@BarackObama	@JoeBiden	@MittRomney	@PaulRyanVP
16.73%	0.87%	7.56%	1.25%

Table 7: Candidate Mention Percentage

5. DISCUSSION

The previous paper presents an explanation of the methods used to construct a dataset of political discourse on Twitter during the 2012 Presidential Election campaign and present a summary of the activity that occurred within the medium. It is imperative when studying technologies to understand the context and atmosphere along with how individuals use the technology and what the findings may mean from a technological and social perspective [8; 10; 17]. Understanding the technology and the syntactical features along with clearly operationalizing them is a necessary aspect of any research in social media [5]. Further, there are a number of ways that have organically emerged to use syntactical features alternatively than the ones that are commonly accepted. For example, there are at least seven ways to retweet a

message [18] and at least one other way to use an at-reply (with a period preceding the @ to broadcast it beyond just the two people and their followers, if the privacy setting is chosen). Clear and appropriate conceptualization can lead to reproducible findings and help readers better understand the topic being studied.

The comprehensive description provides a starting point for understanding further research on the topic and allows for an explicit understanding of the dataset published given numerous word limits in academic venues. As identified in previous papers that detail Twitter collection [1] [5], small choices made when constructing a dataset can greatly alter the findings.

There are numerous socio-technical implications for social media research and society at large that are highlighted by the preceding approach and findings. The study highlights numerous methodological implications that have not been identified in other research that examines social media. One of the implications of this research is the manner in which the construction of a theoretically informed dataset that mixes different types of collection terms allows for a richer analysis that extends beyond just one hashtag or set of keywords. In the context of this study, the collection of keywords, hashtags and at-mentions allows for the temporal analysis of syntactical feature adoption and how this varies over time. These differences highlight interesting characteristics of how individuals construct discourse in a technologically mediated environment.

The aggregation of the dataset that was constantly being refined and expanded as new events emerged also allows for a deeper analysis of the 2012 election while it unfolded. In this study, 68 different terms were used to construct the dataset and these were from different categories (candidates, events, promoted hashtags). This diversity allows for the collection of as near a complete dataset of political discourse as possible that occurred during the studied timeframe.

5.1 Dataset Characteristics

The dataset reflects a significant percentage of individuals who use syntactical features and more specifically the official Twitter handles of the candidates to highlight them in the discourse. Candidate mentions comprise a large amount of the actual discourse and the diverse set of queries used to construct the dataset allow for this to be a material finding in the research as the dataset used for analysis is constructed using a diverse set of queries. Using only the candidate name without the technological feature makes the tracking and identification of the discourse more difficult since it requires the tracking of the keywords instead of just the tracking of the mentions of the handle.

This finding differs from previous political research in which most of the discourse was overtly candidate or party centric [3; 16]. The differences in findings between this study and previous work may be a result of the dataset construction or the fact that the election was more than just about the candidates and also about the issues. Further, the ability to use promoted hashtags and other affiliation hashtags such as #romney or #obama may have allowed individuals to indirectly discuss the candidate or the campaign without using the candidate handle and still participate in the discourse.

The overall syntactical feature distribution of this dataset is similar in some ways to other political Twitter datasets and also differs in some regard. The percentage of retweets (55.10%) is similar to previous political datasets of United States politics [24], but differs from similar analysis on European elections [28]. This could be a result of the different environment and dataset construction, but also could speak to a difference in user population (Americans versus Europeans). This finding suggests that culture or other larger societal factors may have an effect on the activity in the technology. This is an area for further research given the differences.

The percentage of at-replys was higher than other datasets representing a higher adoption of the syntactical feature, albeit for something not necessarily conversational in nature as evidenced by the lack of back and forth activity [12]. As most research on Twitter uses the hashtag for creation of datasets, the percentage of tweets with a hashtag is not comparable (since a dataset with a hashtag as the selection criteria means 100% of tweets will have a hashtag). Therefore, these high level statistics establish a basis for further research into mixed syntactical feature datasets.

5.2 Implications for Social Media and Society

The dataset used for this study has allowed for an examination of how technology specific syntactical feature usage varies by day and event and how these syntactical features can be used by individuals to engage with each other, with candidates and share information. This paper is a foundation for a number of studies and is a contribution to knowledge in the explicit description of conceptualization and description of methods used to construct a complex dataset. These contributions traverse a number of existing literatures from the theoretical, practical, substantive and methodological perspectives. The implications for these findings are broad and extend beyond just politics as more communication is occurring online.

The unique dataset construction allows for more granular approaches to examining the data. In the case of the acute events that were studied, it is possible to examine activity using a variety of units of analysis, such as time or the syntactical features used in the context of an event. In this study, the debates are examined in the context of the whole dataset, but further examination of how a message was retweeted throughout a debate or speech given by the candidates would help to uncover how candidate information diffuses through a network in real time. This type of analysis is possible given the broad collection terms that were used to create the complete dataset. This type of method could also be applied more broadly to the time period in general and understand how individuals respond and share candidate-specific information over the course of an election.

Beyond the scope of this dataset, it is possible to apply this methodological approach and findings to a smaller dataset for a regional or local election and design a study to gather attitudes of voters that use social media to compare their offline characteristics with their online activity. This type of study design would allow for a more thorough understanding of the intent of the online actions and would rival seminal studies on voter behavior and attitudes in complexity and impact [14]. This type of study would also allow for greater insight into how social media is used in the context of political activity in society and would address one of the most significant shortcomings of all social media research – the lack of offline corroboration of activity. This shortcoming limits how findings of technological activity can be used to better understand greater societal activity.

An ongoing examination of how social media is used in the political context is needed. The methodological approaches of dataset construction, collection and analysis need to be documented and shared with the broader community of political and socio-technical researchers. In addition to detailed methodological documentation, researchers need to be willing to share the datasets along with processing and analysis scripts used to arrive at the findings along with clarity surrounding how data

was collected as detailed in chapter three and the appendices included in this document. This transparency will allow for the field to continue to advance and allow for researchers with different perspectives to derive insights from the same datasets. This will narrow the gap that currently exists between those that can collect interesting datasets and those with interesting research questions that can be answered using datasets that they may not have technical capabilities to collect.

This paper hopes to establish a conversation and continue to set a precedent for full disclosure on the methods and decisions used to construct datasets for academic research as small decisions in the collection phase can have material impact on findings. These differences in findings will have a continuing effect on the foundations of social media's effects on society and how individuals and groups are engaging in the numerous technological mediums.

6. ACKNOWLEDGMENTS

This work was supported by NSF grant VOSS-1422982.

7. REFERENCES

[1] Black, A., Mascaro, C., Gallagher, M., and Goggins, S. 2012. TwitterZombie: Architecture for Capturing, Socially Transforming and Analyzing the Twittersphere. ACM Group.

[2] Borah, P. 2014 Facebook Use in the 2012 USA Presidential Campaign. In Social Media in Politics, Springer.

[3] Bruns, A. and Burgess, J. E. 2011. #Ausvotes: how Twitter covered the 2010 Australian federal election. Communication, Politics and Culture. 44, 2, 37-56.

[4] Christensen, C. 2013. Wave-riding and hashtag-jumping: Twitter, minority 'third parties' and the 2012 US elections. Information, Communication & Society. 16, 5, 646-666.

[5] Driscoll, K. and Walker, S. 2014. Big data, big questions working within a black box: Transparency in the collection and production of big twitter data. International Journal of Communication. 8, 20.

[6] Fairclough, I. and Fairclough, N. 2012 Poltical Discourse Analysis: A Method for Advanced Students. Routledge.

[7] Fairclough, N. 1992. Discourse and text: linguistic and intertextual analysis within discourse analysis. Discourse & Society. 3, 2, 193-217.

[8] Goggins, S., Mascaro, C., and Valetto, G. 2012. Group Informatics: A Methodological Approach and Ontology for Understanding Socio-Technical Groups. Journal of American Society for Information Science.

[9] Goggins, S., Galyen, K., and Laffey, J. 2010. Network Analysis of Trace Data for the Support of Group Work: Activity Patterns in a Completely Online Course. ACM Group.

[10] Goggins, S. P., Valetto, G., Blincoe, K., and Mascaro, C. 2012. Creating a model of the Dynamics of Socio-Technical Groups using Electronic Trace Data. UMUAI.

[11] González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., and Moreno, Y. 2014. Assessing the bias in samples of large online networks. Social Networks. 38, 16-27.

[12] Honeycutt, C. and Herring, S. C. 2009. Beyond Microblogging: Conversation and Collaboration via Twitter. Hawaii International Conference on System Sciences. 43.

[13] Hong, S. and Nadler, D. 2012. Which candidates do the public discuss online in an election campaign?: The use of social media by 2012 presidential candidates and its impact on candidate salience. Government Information Quarterly. 29, 4, 455-461.

[14] Huckfeldt, R. and Sprague, J. 1987. Networks in Context: The Social Flow of Information. American Political Science Review. 81, 4.

[15] Johnson, J. 2012. Twitter bites and romney: Examining the rhetorical situation of the 2012 presidential election in 140 characters. Journal of Contemporary Rhetoric. 2, 3/4, 54-64.

[16] Jurgens, P., Jungherr, A., and Schoen, H. 2011. Small Worlds with a Difference: New Gatekeepers and the Filtering of Political Information on Twitter. WebSci.

[17] Karpf, D. 2012. Social science research methods in Internet time. Information, Communication & Society. 15, 5, 639-661.

[18] Kooti, F., Yang, H., Cha, M., Gummadi, K., and Mason, W. A. 2012. The Emergence of Conventions in Online Social Networks.

[19] Kreiss, D. 2014. Seizing the moment: The presidential campaigns' use of Twitter during the 2012 electoral cycle. New Media & Society. 1461444814562445.

[20] Mascaro, C., Agosto, D., and Goggins, S. 2016. One-Sided Conversations: The 2012 Presidential Election on Twitter. International Digital Government Research Conference.

[21] Mascaro, C., Novak, A., and Goggins, S. 2012. The Daily Brew: The Structural Evolution of the Coffee Party on Facebook During the 2010 United States Midterm Election Season. Journal of Information Technology and Politics. 9, 3, 234-253.

[22] Mascaro, C., Black, A., Gallagher, M., and Goggins, S. 2012. The 2012 Wisconsin Gubernatorial Recall Twitter Corpus. Available at SSRN 2159303.

[23] Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. M. 2013. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. Proceedings of ICWSM.

[24] Mustafaraj, E. and Metaxas, P. T. 2011. What Edited Retweets Reveal about Online Political Discourse. Workshop on Analyzing Microtext.

[25] Petrovic, S., Osborne, M., and Lavrenko, V. 2010. The edinburgh twitter corpus. Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, 25-26.

[26] Sharp, A. 2012. Election Night 2012. Twitter.

[27] Thelwall, M. 2015. Evaluating the Comprehensiveness of Twitter Search API Results: A Four Step Method. cybermetrics.info.

[28] Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. ICWSM. 4th.

[29] van Dijk, T. A. 1997. What is Political Discourse Analysis? Belgian Journal of Linguistics. 11, 1, 11-52.

[30] Wang, H., Can, D., Kazemzadeh, A., Bar, F., and Narayanan, S. 2012. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. Proceedings of the ACL 2012 System Demonstrations, 115-120.