Real-Time Filtering for Pulsing Public Opinion in Social Media

Samantha Finn and **Eni Mustafaraj**
Wellesley College
Wellesley, MA 02481
sfinn@wellesley.edu, emustafa@wellesley.edu

Abstract

When analysing social media conversations, in search of the public opinion about an unfolding event that is being discussed in real-time (e.g., presidential debates, major speeches, etc.), it is important to distinguish between two groups of participants: opinion-makers and opinion-holders. To address this problem, we propose a supervised machine-learning approach, which uses inexpensively acquired labeled data from mono-thematic Twitter accounts to learn a binary classifier for the labels "political account" (opinion-makers) and "non-political account" (opinion-holders). While the classifier has a 83% accuracy on individual tweets, when applied to the last 200 tweets from accounts of a set of 1000 Twitter users, it classifies accounts with a 97% accuracy. This high accuracy derives from our decision to incorporate information about classifier probability into the classification. Our work demonstrates that machine learning algorithms can play a critical role in improving the quality of social media analytics and understanding, whose importance is increasing as social media adoption becomes widespread.

Introduction

Real-time event reporting and commenting is a popular activity in social-networking sites such as Facebook and Twitter. However, these two websites approach the networking aspect differently: Facebook is about staying in touch with friends and family, thus, most of the messages are per default private; Twitter is about staying in touch with the rest of the world and per default messages are public. Moreover, with built-in features such as hashtags, retweeting, and trending topics, Twitter makes it very easy for news to be discovered early and then spread quickly through the network. There are several situations in which this real-time diffusion is important: the United States Geological Survey (USGS) has developed a Twitter-based earthquake detection system (Hurs 2010), since in populated regions Twitter reports are faster than USGS alerts. In Mexico, where traditional media are censored by the threat of both government and drug gangs, citizens have turned to Twitter to report and confirm in real-time life-threatening events such as shootings and kidnappings (Monroy-Hernandez 2011).

In addition to breaking news, the most popular real-time conversations relate to televised events: important sportscasts, popular TV shows, or political events. In fact, many political events gather a lot of attention in Twitter: press conferences by President Obama, debates of presidential candidates for the Primary Elections in 2012, or congressional hearings of the Super-Committee.

The increase in Twitter activity during real-time events has not gone unnoticed. For years, traditional media have been trying to answer the inevitable question during presidential debates: who won the debate tonight? A rather simple solution has been to gather an audience of diverse citizens in a room and ask for their opinion. A more expensive solution has been to involve a polling company to perform phone interviews with a carefully selected sample. However, with the wide-spread adoption of social network websites, there is a new source of information: citizens discussing the event live and offering their comments and opinions about what is happening. In 2008, during one of the debates between Barack Obama and John McCain[1], for the first time researchers were able to characterize debate performance moment-by-moment, by analyzing the sentiment of tweets collected during the live debate (Diakopoulos and Shamma 2010). More researchers have followed this path by using the volume of tweets and their sentiment in order to predict events such as political elections (Tumasjan et al. 2010). However, such approaches have an inherent drawback: they treat all Twitter accounts as equal. Knowing who is sending a tweet turns out to be important, because not all accounts belong to humans and even when they do, some of them are actually not "normal" social network users.

Who is tweeting?

With more people using Twitter, the incentive for spammers and other mischiefs to infiltrate the network has been increasing. Research has shown that the click-through ratio for spam URLs incorporated in tweet posts is much higher than in email (Grier et al. 2010). Additionally, there is another kind of spam: Twitter accounts that are used to spread misinformation about political candidates (Metaxas and Musta-

[1] In 2008, Twitter was not yet very popular. Its adoption increased significantly around March 2009 when celebrities such as Oprah started publicizing it in the media.

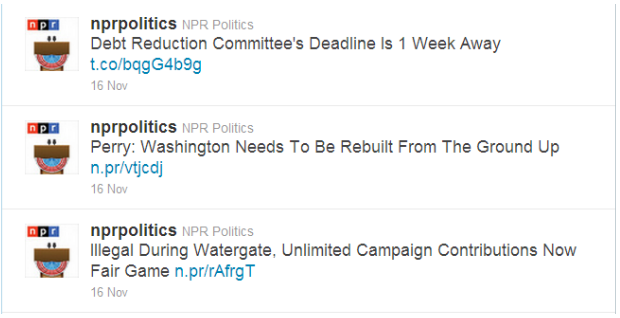
faraj 2010). Although such accounts have a clear adversary agenda, there are other types of accounts on Twitter that do not behave normally:

- broadcaster accounts. Such accounts broadcast headlines non-stop. An example is @greychampion who has posted 600,763 tweets[2]. In this category are also the large number of media organization accounts (CNN, Fox News, NPR, etc.)
- activist accounts. Such accounts belong to political activists (e.g., lobbying groups) who constantly tweet and retweet the same content for political motives.

Because the nature of such accounts is different from normal social network users, who tweet about a variety of topics: their personal life, concerns, thoughts, etc. (Naaman, Boase, and Lai 2010), every study that tries to pulse the public opinion about a certain event based on social media data needs to distinguish between different kinds of Twitter accounts.

The need for filtering

In traditional public opinion polling techniques, careful deliberation goes into choosing a representative sample (Blumenthal 2004). We believe that the same care should be applied when analysing social media signals for capturing public opinion. To this purpose, machine learning techniques can be very helpful. Research has shown that it is possible to accurately infer several demographic features, such as political orientation or ethnicity (Pennacchiotti and Popescu 2011) and distinguish between humans and bots in Twitter (Chu et al. 2010). In this paper, we consider a new problem: distinguishing between political and non-political accounts. A "political account" is defined as a Twitter account that tweets almost exclusively about political issues, while a "non-political account" is a Twitter account that tweets about a broad range of topics and is not primarily focused on politics. It is important to distinguish between these two types of accounts when pulsing public opinion because we want to separate opinion-makers from opinion-holders (the general public). Political accounts belong to politicians, pundits, journalists, political organizations, political lobbyists, etc. They tweet almost exclusively about politics and their primary reason for tweeting is to promote their opinions. By separating the tweets from these two categories of accounts, we will be able to analyze them separately and answer questions such as: what is the opinion of general public, what is the opinion of opinion-makers, or which of the opinion-makers is the most influential (whose opinions spread further in the social network). In this paper, we discuss a supervised machine learning approach that uses automatically acquired labeled data from monothematic Twitter accounts. Testing in 1000 accounts divided evenly among the two classes of political and non-political accounts demonstrates that such inexpensively acquired labeled data are a reliable source of knowledge.



nprpolitics NPR Politics
Debt Reduction Committee's Deadline Is 1 Week Away
t.co/bqgG4b9g
16 Nov

nprpolitics NPR Politics
Perry: Washington Needs To Be Rebuilt From The Ground Up
n.pr/vtjcdj
16 Nov

nprpolitics NPR Politics
Illegal During Watergate, Unlimited Campaign Contributions Now Fair Game n.pr/rAfrgT
16 Nov

Figure 1: Example of political tweets by @nprpolitics, the Twitter account for NPR's political section.

Collecting Data on Twitter

Twitter is a social network and microblogging service that allows its registered users to send and receive messages up to 140 characters known as tweets (or status updates). Twitter is a directed network (unlike Facebook). Links from a user to other users in the network are known as friendship links, the ones from other users to a certain user as following links. When a user logs into Twitter, she gets to see all tweets written by her friends. When she tweets, all her followers see her tweets. According to statistics published in September 2011 (Hachman 2011), Twitter currently has 100 million active users worldwide who write approximately 200 million tweets per day. According to Pew Research, 13% of American adults are on Twitter (Smith 2011), including 82% of the U.S. House and 85% of the Senate members.

In order to build a binary classifier that can distinguish between a political account or non-political Twitter account, we need numerous sample tweets that are considered political or non-political. On Twitter, there are many accounts related to news organizations which tweet exclusively about politics such as @nprpolitics, @foxnewspolitics, and @nypolitics. For an example, refer to the tweets from @nprpolitics shown in Figure 1. In the same way, there are accounts that tweet exclusively about topics such as pop culture, lifestyle, or technology, for example @tvguide, @jezebel, and @mashable.

By using the Twitter REST API[3] we collected the most recent 3200 tweets of 10 accounts, referred to as our tweet-training dataset. The number 3200 is a limit set by the Twitter API. Because different accounts use Twitter differently, the earliest tweets in the collected data fall in different time periods; thus, our collection covers a period of almost three years, as Table 1 indicates. Our scripts collected 11,034 political tweets and 17,369 non-political tweets. As we discuss in the following section, the tweets from these accounts are used to build our classifier for the two classes: political and non-political.

Political vs. Non-Political Accounts

The primary goal of building a classifier is to distinguish between Twitter accounts that are political (we refer to them

[2] As of February 16, 2011.

[3] https://dev.twitter.com/docs/api

political account	date	non-political account	date
nprpolitics	10/10/2010	GlobeWorld	06/05/2009
nypolitics	12/01/2008	jezebel	04/18/2011
politico	07/10/2011	nerve	03/14/2011
foxnewspolitics	02/08/2011	mashable	08/30/2011
politicalticker	08/02/2011	tvguide	07/30/2011

Table 1: Date of the earliest tweet in our training dataset for the political and non-political accounts.

as opinion-makers), and accounts that are non-political (or opinion-holders). In order to test the accuracy of our classifier, we need accounts for which we already know the labels. We used available websites and Twitter suggestions to collect 500 political accounts and 500 non-political accounts. As political accounts, we collected all members of the US Congress (a list can be found at tweetcongress.com), accounts suggested by Twitter in the political category, and a group of political activists identified from our previous research. As non-political accounts, we used lists of famous Twitter users compiled in the web, for example http://listorious.com/Jason˙Pollock/twitter-giants, filtering out any politicians. Then, for each of these 1000 accounts, we collected their most recent 200 tweets, since this can be done with a single Twitter API call per account. This serves as our account-training dataset.

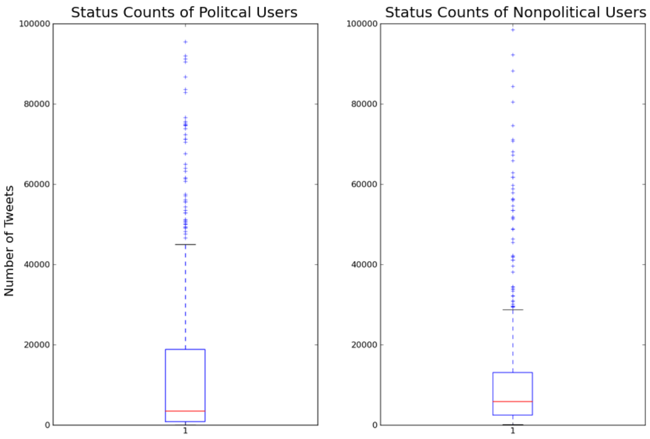


Figure 2: Boxplot showing the number of tweets from the account-training dataset, separated by political and non-political users. In the graph, the status counts are cut off at 100,000, excluding 16 political accounts and 2 non-political accounts. The median value for political account is 4,112 tweets and for non-political accounts 5,873.

The retrieved information contains data that refers to the total number of tweets (statuses count) that the account has tweeted since its creation. To better understand the behavior of the chosen accounts, we plotted the distribution of statuses counts as boxplots (refer to Figure 2). There were 16 political accounts and two non-political accounts that had more than 100,000 tweets, which we removed from the graph in order to not skew the distribution. These accounts are the kind of broadcaster accounts that we would like to filter from public opinion analysis, since they clearly don't represent humans. Because the boxplots show that the two distributions overlap, the number of statuses alone is not a good discriminator between these two types of accounts.

Data Processing

Though a tweet is at most 140 characters, it contains different entities that makes it rich in structure: mentions (for example, @BarackObama), hashtags (for example, #Obama2012), and shortened URLs. Since all these entities would vastly increase the size of a vocabulary, the first step for processing tweets is to remove them from the text. In addition, we also remove stopwords, numbers, punctuation, and lowercase everything. No stemming is performed. An example of such processing is shown here:

```
Join all of us at #Obama2012 in
wishing @VP Biden a happy birthday today:
http://t.co/uKYz3Um7
```

```
join us wishing biden happy birthday today
```

In order to have a balanced training set, we processed an equal number of tweets from each group (11,000 tweets/group). Processed political tweets have an average of 8.7 words, while non-political tweets an average of 9.3 words. A summary of the top ten most frequent words in each group is shown in Table 3. As it is usual with bag-of-words approaches, the received data are not perfect. For example, the meaning of the second most frequent political word *house* is not the usual meaning of the word, but it might refer to two different institutions: the White House or the House of Representatives. Meanwhile the word *us* refers to the abbrevation for the United States and not to the personal pronoun (present in the non-political list of frequent words). While a word-disambiguation step might be useful, we are trying to apply minimal natural language processing, to make the method feasible for real-time analytics.

Classifying Accounts

Research in text classification has shown that a Naive Bayes classifier performs as well as more sophisticated classifiers (Huang, Lu, and Ling 2003). Since Naive Bayes is simpler to train, we used it as our learning algorithm. Table 4 shows the ten most informative features for the political and non-political class, and Table 2 summarizes the result of pairing different political and non-political accounts from our tweet-training dataset in the learning phase. All accuracies are calculated by a 10-fold cross-validation process. It is interesting to notice that some pairings perform much

	Political Accounts				
Nonpolitical Accounts	nypolitics	foxnewspolitics	politico	nprpolitics	politicalticker
globe accounts	0.93	0.82	0.81	0.94	0.87
jezebel	0.95	0.91	0.90	0.92	0.96
mashable	0.98	0.87	0.86	0.96	0.88
nerve	0.98	0.84	0.83	0.92	0.86
tvguide	0.99	0.87	0.86	0.96	0.88

Table 2: Accuracies of classifiers trained on different political and nonpolitical accounts. When combining all the political and nonpolitical accounts, the classifier accuracy is 0.83.

political	frequency	non-political	frequency
obama	1359	new	782
house	745	blog	322
gop	712	video	309
new	526	get	296
senate	498	tv	269
says	479	like	240
us	460	says	237
bill	413	watch	232
debt	401	us	229
budget	387	facebook	217

Table 3: The most frequent words in the political and non-political tweet-training dataset.

political feature	non-political feature
democrats	globe
libya	iphone
boehner	boston
senate	dishing
debt	pics
sen	bases
mitt	doc
committee	season
budget	ipad
congressional	eats

Table 4: The most informative words from the classifier trained on the political and non-political tweet-training dataset.

better than others: for example, the pairing of @tvguide and @nypolitics has a 98% classification accuracy. However, since we want to use our classifier with Twitter accounts that are not mono-thematical, we decided to train the classifier in all the tweet-training tweets. This classifier performs with a 83% accuracy. The accuracies in Table 2 suggest that a multi-class classifier that is able to recognize different topics might perform better and we will address this in our future research.

Two-step classification

Being able to classify tweets as political or non-political is only the first step in classifying Twitter accounts as political or not. Our base assumption is that both these kinds of accounts will have a mix of these two types of tweets, but in different proportions. In order to determine the decision boundary in an optimal way, we followed a two-step classification strategy to classify an account:

1. Classify 200 tweets of each account as political or non-political, using our Naive Bayes classifier trained with labeled tweets.
2. Classify each account as political or non-political, using a one-feature (ratio of political tweets) linear classifier trained on labeled accounts.

The Naive Bayes classifier assigns a probability to every label and in order to improve the accuracy of our classification approach, we decided to use a threshold of 0.9; that is, only tweets classified with a probability of 0.9 or higher were included in the second step of deciding the label for the account. From the 1000 account-training user, 66.7% of all classified tweets passed this threshold. Only 9 accounts had no tweets with a probability of 0.9 or higher and were not included in the second classification step. Of the 991 remaining accounts, an average of 133 tweets per account was used to determine their class.

A simple linear classifier was trained to learn the decision boundary between the two classes, based on the ratio of political tweets to the total number of classified tweets per account. The training was conducted on our 1000 (equally divided) accounts and resulted in a boundary of 22.8%. As the boxplot chart (Figure 3) for the distribution of the ratio values in the two classes shows, that is an appropriate boundary which misclassifies only some outliers from each class, as also summarized in the confusion matrix in Table 5.

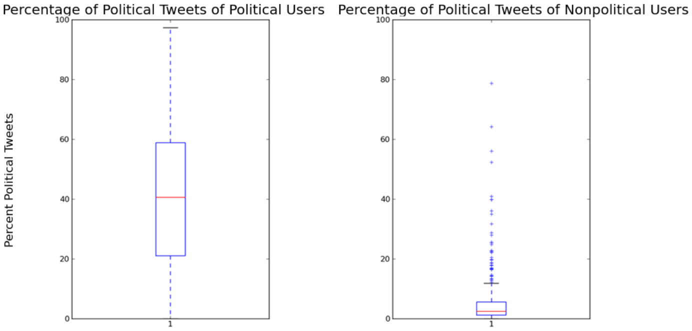


Figure 3: Boxplots comparing the ratio of political tweets for political and non-political accounts. Only tweets classified with probability > 0.9 were included.

	Predicted Class	
Actual Class	Political	Non-political
Political	488	8
Non-political	22	473

Table 5: Confusion matrix for classification of Twitter accounts. It is based on the described two-step classification process. Nine accounts didn't have any tweets with a classification probability > 0.9 and thus couldn't be included.

Discussion

Hashtags as Topical Markers

To better assess the classification process, we ranked labeled tweets based on classifier probability, and manually verified the labels of 110 tweets evenly divided into two groups of political and non-political tweets. The classifier did very well with political tweets (error rate 5.5%), but did poorly with non-political tweets (error rate 21.8%). We inspected all the tweets that were incorrectly labeled as non-political and we discovered that most of them contained the hashtag #AskObama. These tweets were part of a collection with tweets directed to President Obama during his Twitter town hall meeting[4]. Because during our pre-processing we had excluded hashtags (since often they are not real words and many of the accounts in our tweet-training data do not use hashtags), the text of the tweet without the hashtag could easily be interpreted as non-political. For some examples, refer to tweets in Table 6.

Since our classifier uses a bag-of-words model that doesn't capture the meaning of text, the fact that hashtags could potentially contain the kind of useful information that allows to discriminate between tweets makes it important to be able to deal with them properly. What makes hashtags difficult to deal with is that the majority of them are ephemeral: they are created for a single event and then cease to exist. However, since Twitter users tend to make use of many hashtags in their tweets, it might be possible to exploit co-occurrence of hashtags to infer some kind of meaning. Thus, for all tweets in our corpus that contained the hashtag #AskObama, we counted the frequency of co-occurring hashtags and the top ten hashtags are shown in Table 7. The first observation from this table is that two of the top three hashtags are so-called topical hashtags that are used to mark tweets belonging to certain political groups: #99ers refers to the group of American citizens for which the 99 weeks of unemployment benefits have expired, #p2 refers to the progressive movement. Research has shown that political hashtags, in contrast to other kinds of hashtags in Twitter, have longevity (Romero, Meeder, and Kleinberg 2011). Furthermore, politically engaged users adopt them consistently in their tweets (Conover et al. 2011). These features of political hashtags make them a strong candidate for use in semi-supervised learners, where the established hashtags are used as seeds which then allow the learner to infer the nature of other co-occurring hashtags. The second observation is that most of the other hashtags are also political, though they are the kind of event-driven hashtags that correspond to discussion of the news related to phenomena such as WikiLeaks or the Anti-Security movement. Thus, by feeding to a system a list of known and persistent political hashtags, it might be possible to discover ephemeral political hashtags which would improve the quality of classification in those cases when the bag-of-words approach is not robust. This is an interesting area of future work that we are pursuing.

hashtag	count
#99ers	203
#noagenda	188
#p2	183
#anonymous	171
#wikileaks	164
#cnndebate	160
#antisec	159
#brazilnocorrupt	158
#lulzsec	158
#poker	157

Table 7: Top ten hashtags in the #askobama data set (excluding #askobama with a count of 6985).

Abnormal Activities

As the results of classification shown in Table 5 indicated, the classifier was less successful in identifying non-political accounts (error 4.4% versus 1.6% for political accounts). In fact, these errors can be clearly seen in Figure 3 as outliers in the boxplot for non-political accounts. We manually inspected the tweets of one non-political account that showed the largest error. Out of 200 tweets, 165 were classified with a probability higher than 0.9. 130 tweets out of these remaining 165 were classified as political, and therefore the account itself was classified as political. The account was @biz, which belongs to Biz Stone (one of the Twitter co-founders). On November 8, 2011, he had sent more than one hundred tweets to different Twitter users to urge them to vote in the San Francisco election for Mayor. Tweets were very similar and all sent automatically from the website Votizen.com. Here are some examples:

```
@fledgling Will you endorse @EdLeeforMayor
for Mayor on @Votizen? http://t.co/dsZgUo4G
```

```
@epicentercafe Will you endorse
@EdLeeforMayor for Mayor on @Votizen?
http://t.co/dsZgUo4G
```

Normally, the Twitter account of @biz is not a political account. However, because he engaged in a high-volume political activity in the time-period of the data collection, he was rightfully labeled as a political account. But this raises many interesting questions. Should we base the classification on the last 200 tweets (which can be quickly retrieved with a single API request) or should we retrieve all data we can from an account (not more than 3200 tweets) and then choose random tweets for the classification. That will have the disadvantage of being slower (for real-time filtering), but might offer better accuracy. Clearly though, being able to

[4] http://askobama.twitter.com/

Nr	Tweet text	True label	Classifier
1	@townhall #AskObama do you want your girls to live in a world where they can't breath the air - eat the fish or swim in water?	p	non-p
2	@townhall #AskObama greed causes poverty - lets look @ the wealthy & how they earned it 2 discover the real crimes going on start with oprah	p	non-p
3	But not Afghanistan RT @badler: I like how Obama pronounces words like Pakistan and Taliban correctly	non-p	p

Table 6: Tweets mislabeled by our classifier. Notice how the 1st and 2nd tweet contain many words that usually appear in non-political context, such as fish, swim, discover, greed, Oprah. The 3rd tweet is a non-political tweet because is commenting Obama's pronunciation of political words, which explains why the classifier has classified as a political tweet.

discover such abnormal behavior is also a very useful feature, and we will be investigating both sides of this problem in our future work.

Conclusion

One of the traditional problems in supervised machine learning is how to acquire training data. Many current researchers are using Mechanical Turk as a source to get qualitative labels in an inexpensive way, though the time and effort to fight abusive behavior in Mechanical Turk has its own cost. Depending on the nature of the classification problem, the social web might offer alternative ways of obtaining training data. As demonstrated in this paper, we were able to easily acquire labeled instances by collecting data from particular Twitter accounts that have mono-thematical nature (such as @nprpolitics, @foxnewspolitics, @GlobeSox, @BostonAE). A Naive Bayes classifier built upon two sets of such labeled tweets (political tweets and non-political tweets) demonstrated a 83% accuracy. Though at the level of single tweets such a classifier is not very reliable, by using the probability information for each label and making use of a large number of tweets for every account, we were able to classify Twitter accounts as political or non-political with the high accuracy of 97%.

Many interesting questions remain. How reliably can the classifier work with Twitter accounts that have far less than 200 tweets? How often do we need to update the classifier in order to incorporate new vocabulary? How can we make use of the implicit knowledge of hashtags? With more and more people using social media for everyday communication, the need for sense-making tools that allow insights in such communication will only grow. Machine learning solutions will be at the heart of such tools, but we need to be prepared in identifying what kind of problems machine learning algorithms can solve reliably and how they can be solved.

References

Blumenthal, M. 2004. The why and how of likely voters. http://www.mysterypollster.com/main/2004/10/the˙why˙how˙of˙.html.

Chu, Z.; Gianvecchio, S.; Wang, H.; and Jajodia, S. 2010. Who is tweeting on twitter: Human, bot, or cyborg? In *In Proc. of ACSAC '10*. ACM.

Conover, M.; Ratkiewicz, J.; Francisco, M.; Gonalves, B.; Flammini, A.; ; and Menczer, F. 2011. Political polarization on twitter. In *In Proc. of ICWSM '11*. AAAI Press.

Diakopoulos, N., and Shamma, D. A. 2010. Characterizing debate performance via aggregated twitter sentiment. Conference on Human Factors in Computing Systems.

Grier, C.; Thomas, K.; Paxson, V.; and Zhang, M. 2010. *@spam: The Underground on 140 characters or less*. ACM. 27–37.

Hachman, M. 2011. Twitter continues to soar in popularity, sites numbers reveal. http://www.pcmag.com/article2/0,2817,2392658,00.asp.

Huang, J.; Lu, J.; and Ling, C. X. 2003. Comparing naive bayes, decision trees, and svm with auc and accuracy. In *Proceedings of the Third IEEE ICDM03*.

Hurs, T. 2010. Usgs develops twitter-based earthquake detection system. http://bit.ly/7tEndi.

Metaxas, P. T., and Mustafaraj, E. 2010. From obscurity to prominence in minutes: Political speech and real-time search. In *Web Science 2010*.

Monroy-Hernandez, A. 2011. Shouting fire in a crowded hashtag: Narco censorship & "twitteroristas" in mexico's drug wars. http://rww.to/n5nTbT.

Naaman, M.; Boase, J.; and Lai, C. 2010. Is it really about me? message content in social awareness streams. In *In Proc. of CSCW 2010*.

Pennacchiotti, M., and Popescu, A.-M. 2011. A machine learning approach to twitter user classification. In *In Proc. of ICWSM '11*. AAAI Press.

Romero, D.; Meeder, B.; and Kleinberg, J. 2011. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *Proceedings of WWW Conference*.

Smith, A. 2011. Twitter update 2011. http://pewinternet.org/Reports/2011/Twitter-Update-2011.aspx.

Tumasjan, A.; Sprenger, T.; Sandner, P. G.; and Welpe, I. M. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proc. of 4th ICWSM*, 178–185. AAAI Press.