

A Comparison of Social Network Analysis and Text Regression using Twitter Data

Drew Conway and John Myles White

August 15, 2011

1 Abstract

Within the social sciences interest in statistical methods for the automatic analysis of text has grown considerably in recent years. Specifically, within the field of Political Science there has been continued interest in spatial models of political ideology based on text analysis (Grimmer, 2010; Monroe et al., 2008; Laver et al., 2003). Moreover, interest in how the structure of social networks affects political attitudes and outcomes has been both growing (Siegel, 2009; Hafner-Burton et al., 2009) and controversial (Fowler and Christakis, 2010; Lyons, 2010). In an attempt to bridge the gap between these two research areas, in the work presented here we have collected a large sample of Twitter data related to Congress. Twitter provides a rich data set containing both text and social network information about its members. Here we compare the usefulness of text analysis and social network analysis for predicting the political ideology of Twitter users—both members of Congress and the surrounding network of Twitter user. To do so, we estimate a model of political ideology using text as inputs: we fit the model to data from Representatives and test it against members of the Senate. Simultaneously, we estimate a model of political ideology using only social network information as inputs. Again, we fit the model to data from Representatives and test it against members of the Senate.

In this preliminary study we find that each method provides novel insight into the ideological spectrum of the U.S. Congress. The text analysis illustrates the lexical context the manifest this spectrum. On the other hand, the network model provides an example of how the position of specific broadcaster nodes from a network can imply political ideology among receiver nodes.

2 Methods

Using data from the Sunlight Foundation on which members of the U.S. Congress maintain a presence on Twitter, we have identified 67 Senators and 313 Representatives.¹ From these members of Congress, we have generated a data set of nearly one-hundred thousand tweets from members of the U.S. Congress between the period of November 11, 2007 and August 15, 2011. The data were harvested using a spider written in Python that makes calls to Twitter’s API on a regular basis.²

¹The Sunlight Foundation provides this information via its API at http://services.sunlightlabs.com/docs/Sunlight_Congress_API/.

²The code used to generate this corpus, and all other data related to this project, can be inspected and downloaded here: <https://github.com/drewconway/Twitter-Congress>. Readers should note, however, that due to Twitter’s data

Each tweet in the data set contains the following information: full name of member of Congress, title (Senator or Representative), party affiliation, home state, gender, Twitter user name, text of tweet, date and time the tweet was published, and the unique ID number for the tweet within Twitter’s database.

In addition to this source of text data, we have used Google’s SocialGraph API to generate the directed social graph for Twitter.³ The full graph contains 418,941 nodes with 926,385 edges. The largest weakly connected component of this graph contains 418,941 nodes and 926,385; a trivial loss of structure given the scale of this network, and thus this main component is the focus of this research.

Table 1: Summary statistics of Main Component of Compressional Twitter Network

Nodes	Edges	Density	Mean Degree
418,941	926,385	2.21	$5.28e^{-6}$

These data provide a useful testbed for comparing methods measuring political ideology. Because we can use ideal points measured for all of the members of Congress based on roll call data (Jackman, 2001), it is possible to test our predictions rigorously. And the bicameral nature of the Congress provides an obvious mechanism for testing a predictive model on held out data: we fit our models to data from the House and then test the models on data from the Senate.

Given both text data and social network data, there are two obvious models that we can fit: a text regression model in which the word counts for each tweet are used to predict the ideal point of the tweeting member of Congress; and, a social network model in which political views propagate out through the social network with a given rate of absorption and decay at each node.

By comparing the RMSE of both models on data from the Senate after fitting the models to the House, we can determine the viability of both analytic methods.

2.1 Text Regression

The 96,829 tweets in our data set can be treated as separate observations; for each of these tweets, we observe the number of occurrences of any of the words in our corpus of tweets. Because many words occur only once, we remove the terms that occur in less than N documents. This is considerable reduction in the number of variables we have to work with: we begin with XXX terms and, after pruning, have only YYY terms. Given these measured word counts, we attempt to predict the ideal point of the member of Congress that wrote each tweet.

We fit the model using standard convex optimization techniques: this is implemented in the R package, `glmnet`, which we have used for our analyses. With this, we fit several model variants, including models that use a log transform of the word counts and two Lasso models that use only the hashtags and mentions from the tweets. Even these substantially impoverished models (there are only N hashtags and M mentions) outperform the baseline model under held-out testing.

policies the raw data set of tweets could not be shared. Also, due to the moving window of data available in Twitter’s search API any replication of the data using the code available here may not match exactly what is used in this following analysis.

³We have used Google’s social graph service rather than Twitter’s because our need to build the full social graph of all members of Congress exceeded the limits provided by the Twitter API, but was within Google’s.

2.2 Network Model

To model political ideology on the Congressional network we construct a simple model of transmission using an exponential decay. In this case, we use the Jackman scores as assumed ideal points for one set of members of Congress, i.e., either Representative or Senators. These scores are then “broadcast” over the network, and each node then absorbs this score at an exponential rate of decay based on geodesic distance from the broadcasting node. Equation 1 below describes the form of this model.

$$\hat{\pi}_v = \sum_{i=1}^N \pi_i^{-k} \tag{1}$$

This equation states that the estimated ideology of some node, $\hat{\pi}_v$, is equal to the sum of all broadcast ideologies for some set N of nodes within the network. In the case of the Congressional Twitter network, the set N will either be the set of Representative or Senators. These ideological broadcasts decay at an exponential rate given the geodesic distance k of v from i . Any exponential rate of decay for information transmission has been proposed in the literature (Wu et al., 2004), and given the scale of the network this simple additive model is preferable for computational tractability.

3 Results

The results from the text regressions are described in Table 2 below. (John, add a sentence or two describing this table)

Table 2: Model Comparison for Text Regression Variants

Model	RMSE	R^2
Baseline	1.062	0.00000
Lasso	0.9729	0.08390
Log Lasso	0.9731	0.08371
Ridge	0.9771	0.07994
Log Ridge	0.9774	0.07966
Hashtags	1.037	0.02354
Mentions	1.058	0.00377

(Drew, will try to add results from network model, but likely will not finish before deadline. In that case, add a line that results from network model forthcoming, but will be complete by WIN.)

4 Discussion

We have provided a proof of concept example of the viability of using either social network analysis or text analysis to mine Twitter for insight into the political culture of the U.S. We find that XXX outperforms YYY. All of the methods perform better under held-out data model assessment than the baseline model which predicts the mean ideal point for all members of the Senate.

Future work will need to further improve on the methods we have used and to attempt to combine both social network analysis and text analysis simultaneously.

References

- Fowler, J. H. and N. A. Christakis (2010). Cooperative behavior cascades in human social networks. *Proceedings of the National Academy of Sciences* 107(12), 5334–5338.
- Grimmer, J. (2010). A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis* 18(1), 1–35.
- Hafner-Burton, E. M., M. Kahler, and A. H. Montgomery (2009). Network analysis for international relations. *International Organization* 63(03), 559–592.
- Jackman, S. (2001). Multidimensional analysis of roll call data via bayesian simulation: Identification, estimation, inference, and model checking. *Political Analysis* 9(3), 227–241.
- Laver, M., K. Benoit, and J. Garry (2003). Extracting policy positions from political texts using words as data. *American Political Science Review* 97(02), 311–331.
- Lyons, R. (2010, July). The Spread of Evidence-Poor Medicine via Flawed Social-Network Analysis. *ArXiv e-prints*.
- Monroe, B. L., M. P. Colaresi, and K. M. Quinn (2008). Fightin’ words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis* 16(4), 372–403.
- Siegel, D. A. (2009). Social networks and collective action. *American Journal of Political Science* 53(1), pp. 122–138.
- Wu, F., B. A. Huberman, L. A. Adamic, and J. R. Tyler (2004). Information flow in social groups. *Physica A: Statistical and Theoretical Physics* 337(1-2), 327 – 335.