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

Drew Conway and John Myles White

September 8, 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 using text analysis to extend well-known spatial models of political ideology (**???**). Moreover, interest in how the structure of social networks affects political attitudes and outcomes has been both growing (**??**) and controversial (**??**). In an attempt to bridge the gap between these two research areas we have collected a large sample of Twitter data related to the U.S. 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 — a method that is, in principle, applicable to both members of Congress (for whom roll call data and precise spatial estimates of political ideology already exist) and to the surrounding network of Twitter users (for whom precise estimates of political ideology do not exist). To compare text analysis methods with social network analysis, we fit a model of political ideology using only text as inputs to data from Representatives and test the model against data from members of the Senate. We also estimate a model of political ideology using only social network information as inputs; again we fit this model to data from Representatives and test it against members of the Senate.

2 Methods

Using data from the Sunlight Foundation, we have identified 67 Senators and 313 Representatives who maintain a presence on Twitter.[1] For these members of Congress, we have generated a data set of nearly one-hundred thousand tweets posted to Twitter 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] Each tweet in the data set contains the following information:

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

[2] The 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 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.

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 the members of Congress using Twitter.[3] Summary statistics for this graph are shown in Table 1.

Table 1: Summary Statistics for Main Component of Congressional Twitter Network

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

Taken together, the text and social network data set provides a useful testbed for comparing methods for measuring political ideology. Because we can use ideal points measured for all of the members of Congress based on roll call data (**?**), 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 96,829 separate observations; for each of these tweets, we observe the number of occurrences of any of the 40,000+ words in our corpus of tweets. Because many words occur only once, we have removed the terms that occur in less than 5 documents. After pruning, we have only 1,361 terms. Given the measured word counts as a predictor matrix, we attempt to predict the ideal point of the member of Congress that wrote each tweet in a standard OLS regression with either L1 and L2 regularization of the fitted coefficients. Additionally, we perform this regression using either (a) only the hashtags mentioned in tweets or (b) only the mentions of other Twitter users occurring in our tweets. Results from all four regressions are reported below.

2.2 Network Model

To model political ideology on the Congressional network we construct a simple model of transmission using exponential decay. 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 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.

[3]We 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.

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

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. An exponential rate of decay for information transmission has been proposed in the literature (**?**), and, given the scale of the network, this simple additive model is preferable for computational tractability.

3 Results

The results from the text regression analyses are shown in Table 2 below. The baseline model uses only the average ideal point to make predictions for Senators, while the L1 and L2 regularized regression models use coefficients for all 1,361 terms in our pruned corpus. The Hashtags Only model is an L1 regularized regression model using a subset of 79 of the predictors from the full L1 model; similarly, the Mentions Only model is an L1 regularized regression model using only 27 of the predictors from the full L1 model.

Table 2: Model Comparison for Text Regression Variants

Model	RMSE	R^2
Baseline Model	1.062	0.00000
L1 Regularization	0.9729	0.08390
L2 Regularization	0.9771	0.07994
Hashtags Only	1.037	0.02354
Mentions Only	1.058	0.00377

Results from the network model are currently being computed and will be reported in the final submitted abstract for the conference.

4 Discussion

Our work provides a proof of concept demonstration of viability of using both social network and text derived from Twitter to model the political culture of the U.S. Future work will need to improve on the fine details of the methods we have employed. Another promising research project is to construct statistical models that employ social network and text data simultaneously as a conjugate set of constraints for the estimation of spatial models of political ideology.