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...

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

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

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

Network	Nodes	Edges	Diameter	Density	Mean Degree
Main Component	418,941	926,385	???	2.21	$5.28e^{-6}$
Representatives					
Senators					

15, 2011.[2] The data were harvested using a spider written in Python that makes calls to Twitter's API on a regular basis.[3] 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.[4] 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 thus this main component is the focus of this research.

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 (**?**), it is possibly 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: (1) 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 (2) a social network model in which political views propagate out through the social network with 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] Due to variation in the volume of each member of Congress' Twitter stream, the temporal coverage among all members is not uniform.

[3] 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.

[4] 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.

2.1 Text Regression

The 92,382 (FIX) tweets in our data set can be treated as 92,382 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.

Despite the considerable pruning we perform before model fitting begins, fitting more than a thousand parameters is likely to induce considerable over-fitting. For that reason, we employ regularized regression methods. We models using both the Lasso and ridge regression (REFS): these amount to penalization of the L1 and L2 norms of the coefficients in the fitted models.

Such regularization imposes a tradeoff between the degree of penalization and the model's prediction error: the variable governing this tradeoff is typically denoted λ in homage to the Lagrange multipliers first used to formulate ridge regression. This value must be set by the modeler somehow: to do so in a principled way, we fit λ using repeated random subsampling of the data from the House. The results of this resampling operation is to find that the optimal value of λ under held-out test assessment of performance is 0.0001. (This value is conditional on the subset of values we tested, which included XXX to YYY.)

After determining the optimal value for λ, we simply 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.

3 Results

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.

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

Table 3: Top 5 Republican and Democratic Terms from Lasso

Term	Weight
#jobs	1.0176
grand	0.8474
#reins	0.6211
#p2	0.6128
obamacare.	0.6039
hayworth	-1.1515
#debt	-1.1711
#askobama	-1.1780
delegation	-1.1994
#libya	-1.5589

Table 4: Top 5 Republican and Democratic Hashtags from Lasso

Term	Weight
#jobs	1.0288
#p2	0.6820
#dadt	0.6780
#reins	0.6307
#tx17	0.6292
#askobama	-1.2113
#pa11	-1.2181
#debt	-1.2209
#healthcare	-1.3877
#libya	-1.6383

Table 5: Top 5 Republican and Democratic Hashtags from Lasso

Term	Weight
@speakerboehner	0.7664
@foxbusiness	0.7126
@foxnews	0.6955
@gopwhip	0.6776
@thehill	0.6505
@cspan	-0.0743
@wsj	-0.1875
@rephensarling	-0.2954
@natresources	-0.5474
@barackobama	-1.1266

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.

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.