Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media

A Penny for your Tweets: Campaign Contributions and Capitol Hill Microblogs

Tae Yano Dani Yogatama Noah A. Smith
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{*taey,dyogatama,nasmith*}@*cs.cmu.edu*

Abstract

Who influences a politician's public statements? In this paper, we explore one plausible explanation: that financial incentives from campaign contributors affect what politicians say. Based on this idea, we design text-driven models for campaign contribution profile prediction. Using a large corpus of public microblog messages by members of the U.S. Congress, we find evidence for such an association, at the level of contributing industries. We find complementary strengths in a simple model (which has better predictive accuracy) and a more complex model (which gives a more intuitive, human-interpretable explanation).

Introduction

In the 2012 U.S. general presidential election, each candidate's campaign spent more than one billion dollars. There is no question that the American political landscape is shaped by the vast amount of money spent by campaigns. To what extent can we measure or characterize the precise nature of campaign contributors' influence on elected officials? In this paper, we consider a new way of measuring such influence by exploring the connection between the campaign contributions a member of Congress receives and his or her public microblog messages.

Microblogs, especially Twitter, have become an integral part of political campaigns, public outreach by politicians, and political discourse among citizens. Automatic analysis of microblog text has the potential to transform our understanding of public opinion (O'Connor et al. 2010), communication between elected officials and their constituents (Golbeck, Grimes, and Rogers 2010), and information flow in society more generally (Lerman and Ghosh 2010). Here we use probabilistic modeling to infer associations between campaign contributions, as made available by the Federal Election Committee, and the text of tweets from members of Congress.

We begin with the assumption that the extent to which campaign contributions influence politicians should be measurable in the *predictability* of those contributions, given the text. Further, judicious use of latent variables can help reveal linguistic cues associated with contributions from specific industries and interest groups.

Data

Our dataset consists of two parts: messages (tweets) from the accounts officially associated with members of Congress (MCs) and 2012 electoral campaign contributions.

Tweets from Capitol Hill

During the period from May 15–October 31, 2012, we collected through Twitter's search API publicly available tweets posted by Twitter handles officially associated with MCs. These handles were collected from Tweet Congress (http://www.tweetcongress.org). We manually filtered from this set MCs who were not seeking reelection in 2012. Although we do not know who authored any of these tweets, we assume that they are, for the most part, rationally and carefully crafted by the MC's staff. Golbeck et al. (2010) manually coded a large corpus of MC tweets and found the majority of messages to be public relations and promotion, not personal. Our (less systematic) analysis of the data leads to a conclusion consistent with their finding.

Each tweet was lightly preprocessed. Hashtags and at-mentions were retained; URLs, non-alphabetic strings, and 134 common stopwords were not. Downcasing was applied, and regular expressions were used to normalize some segmentation and lengthening variation. Finally, words occurring less than 10 times in the corpus were removed, resulting in a vocabulary of 19,233 word types, and an average tweet length of 9 word tokens. More details follow:

	# MCs	# tweets	# words
Republicans	249	166,520	1,574,636
Democrats	189	98,661	949,862
Independents	2	818	7,312
Total	440	265,999	2,531,810

Below, c will always index MCs. We let $\boldsymbol{w}_c$ be the complete collection of word tokens tweeted by c; $\boldsymbol{w}_{c,t}$ is the tth message by c, and $w_{c,t,n}$ is the nth word in that message.

Electoral Campaign Contributions

For each of the MCs in our tweet dataset, we collected 2012 general election campaign contribution data from the publicly available database maintained by the Center for Responsible Politics (CRP; http://www.opensecrets.org). These data were originally released by the Federal Election

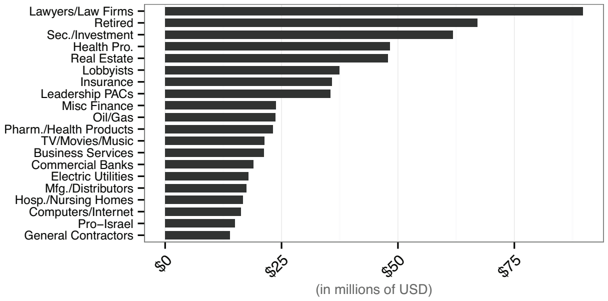

Figure 1: Top 20 contributor industries for the 2012 Congressional Elections (out of 91 in our data); statistics are as of the end of September 2012. Total spending is $1B, mean per industry is $11M, median $6.3M, s.d. $15M.

Commission; CRP performs manual disambiguation and aggregation. Contributions are aggregated by industries and other interest groups defined by CRP. (Hereafter we use the term "industry" to refer to both types of groups.) 91 categories appear in the 2012 data; see Table 1 for the total contributions of the top 20 industries. The variation across industries is very large; lawyers and law firms account for 8% of the total, and the top ten account for 46%.

For each MC, we convert absolute amounts to fractions of the total amount received in contributions. This transformation is meant to help control for variation in spending levels across campaigns, which is large (mean $2.28M, s.d. $2.78M). Fixing the $I = 91$ industries, we let $\boldsymbol{\pi}_c \in \mathbb{R}^I$ denote the **contribution profile** for MC c, where $\sum_{i=1}^{I} \pi_{c,i} = 1$.

Contribution Profile Prediction

To the extent that financial incentives from campaign contributors affect what politicians say, we should be able to infer a MC c's campaign profile $\boldsymbol{\pi}_c$ from his or her tweets $\boldsymbol{w}_c$. We therefore formulate a prediction task to drive our data analysis: given a collection of words $\boldsymbol{w}_c$, how accurately can $\boldsymbol{\pi}_c$ be recovered? We formalize this as a problem of *ranking* industries for an MC and use standard rank comparison scores (Kendall's τ and mean average precision) to compare solutions. Of additional interest for any text-driven prediction model is the interpretability of the learned model, which we consider in qualitative discussion.

Of course, many factors beyond campaign contributions are likely to influence what politicians say, so perfect performance is not expected. In future work, to the extent that other factors can be encoded as evidence, our modeling framework is extendable for their inclusion.

Probabilistic Models

Our approach to the prediction task is to describe a probabilistic model over tweets and campaign profiles, $p(\boldsymbol{W}, \boldsymbol{\Pi})$. This approach requires us to state our assumptions about how the data are generated, and in particular how the two random variables (text $\boldsymbol{W}$ and campaign profile $\boldsymbol{\Pi}$) relate to each other. We consider two different models. In each case, the approach is to reason inductively; we estimate the model parameters from a pool of training examples, and then estimate predictive performance on a held-out test set, discussed in the experiments section.

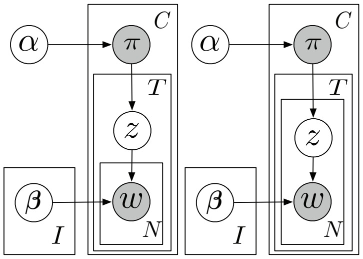

Figure 2: Graphical model representations of our two models, SIPT (left) and MIPT (right). The difference is that SIPT assigns a single industry to each tweet, while MIPT assigns an industry to each *word*.

Single Industry Per Tweet

In our first model, "single industry per tweet" (SIPT), we assume that each tweet is influenced by only one industry. In the first model, the generative story, for each MC c, is:

- Draw a contribution profile $\boldsymbol{\pi}_c \sim \text{Dirichlet}(\boldsymbol{\alpha})$.
- For each tweet by c, indexed by t:
  - Draw an industry $z_{c,t} \sim \text{Multinomial}(\boldsymbol{\pi}_c)$.
  - For each word in the tweet, indexed by n, draw $w_{c,t,n} \sim \text{Multinomial}(\boldsymbol{\beta}_{z_{c,t}})$.

Figure 2 (left) depicts SIPT as a graphical model.

Multiple Industries Per Tweet

Our second model, "multiple industries per tweet" (MIPT) assumes that each tweet is influenced by a mixture of industries, with each word being selected from a different industry. The generative story is, for each MC c, is:

- Draw a contribution profile $\boldsymbol{\pi}_c \sim \text{Dirichlet}(\boldsymbol{\alpha})$.
- For each tweet by c, indexed by t:
  - For each word in the tweet, indexed by n:
    - ∗ Draw an industry $z_{c,t,n} \sim \text{Multinomial}(\boldsymbol{\pi}_c)$.
    - ∗ Draw $w_{c,t,n} \sim \text{Multinomial}(\boldsymbol{\beta}_{z_{c,t,n}})$.

Figure 2 (right) shows the graphical representation of MIPT.

It is worth noting how these two models relate to some familiar probabilistic models for text. SIPT is similar to naïve Bayes classifiers, a model used in text categorization (Nigam et al. 1999). Instead of treating the label of a message as observable, we only observe the MC-level proportions $\boldsymbol{\pi}_c$.

MIPT is similar to latent Dirichlet allocation (Blei, Ng, and Jordan 2003), a model used to infer latent topics in text collections. The topics in LDA are analogous to our industries. The difference is that LDA learns from documents whose associations to topics are completely unknown, so that each $\boldsymbol{\pi}_c$ ($\boldsymbol{\theta}$ in standard LDA notation) is latent. Here, the proportions are observed.

In both cases, the prediction and learning algorithms required are somewhat different from the classic models.

Prediction

Given a new MC c (not included in the training data), we wish to predict $\boldsymbol{\pi}_c$ from messages $\boldsymbol{w}_c$. During prediction, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are fixed. For both models, exactly solving for $\boldsymbol{\pi}_c$

Model	τ	MAP@5	MAP@10	MAP@15
fixed pred.	.49	.52	.60	.65
log. reg.	.49	.41	.49	.55
SIPT	**.55**	**.58**	**.66**	**.70**
MIPT	.44	.42	.51	.57

Table 1: Experimental results. τ is Kendall's τ statistic, and MAP@k is mean average precision at relevance cutoff k.

given the parameters and $\boldsymbol{w}_c$, and summing out all possibilities for $\boldsymbol{z}_c$, is intractable. For SIPT, we apply a single round of message passing, calculating each $z_{c,t}$ based on $\boldsymbol{\beta}$, then $\boldsymbol{\pi}_c$.[1] For MIPT, which involves a more complex latent variable space, we apply mean field variational inference, an approximate technique widely used in Bayesian modeling (Wainwright and Jordan 2008). The details are omitted for space; the algorithm alternates between estimating posteriors over $\boldsymbol{z}_c$ and over $\boldsymbol{\pi}_c$.

Learning by Parameter Estimation

During learning, for a collection of MCs c, $\boldsymbol{\pi}_c$ is observed along with words $\boldsymbol{w}_c$, and the goal is to maximize likelihood with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Because $\boldsymbol{\pi}_c$ is observed, we can estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ separately.[2]

For $\boldsymbol{\alpha}$, we seek the maximum likelihood estimate of the Dirichlet parameters given $\{\boldsymbol{\pi}_c\}_c$. There is no closed-form solution for the MLE, so we apply a well-known fixed-point iterative method (Minka 2000).

For SIPT, we learn $\boldsymbol{\beta}$ using a single round of message passing, calculating each $z_{c,t}$ based on $\boldsymbol{\pi}_c$, then maximizing $\boldsymbol{\beta}$. For MIPT, our algorithm is quite similar to the learning algorithm for LDA given by Blei et al. (2003), but without having to estimate posteriors over tweet-level proportions (since they are observed). As in standard LDA, there is a closed-form solution for maximization over $\boldsymbol{\beta}$. We put a symmetric Dirichlet prior on $\boldsymbol{\beta}$ ("add-one" smoothing).

Experiments

Our experiments are based on 44-fold cross-validation. In each of 44 trials, we held out five (distinct) MCs with tweet volume between the 25th and 75th percentiles. The remainder of the MCs' data were used to estimate parameters, and then predictions were made for the held-out MCs. Each prediction is a ranked list of industries.

We include a fixed prediction that ignores tweet content. This is the ranking of industries based on the training data for the fold. We also include a multinomial logistic regression-inspired discriminative model as a simpler machine learning technique. This model is trained using $\boldsymbol{\pi}_c$ to define a fractional assignment of MC c across industries.[3]

Industry	Associated Terms
Computers and Internet	#sopa, internet, sopa, rights, tech, ag, property, judiciary, holder, oppose, federal, open
Defense (Electronics)	security, border, homeland, cyber, subcommittee, hearing, defense, air, nuclear, briefing, mike, turner, military
Defense (Aerospace)	defense, afghanistan, armed, appropriations, services, inouye, subcommittee, committee, budget, secretary, military, fort
Agr. Services/Products	harkin, iowa, farm, announces, farmers, qs, ag, nebraska, drought, webcast, nelson, agriculture
Agr. Tobacco	nc, #ncsen, burr, #nc7, hours, carolina, office, schedule, #ncpol, north, county, staff
Fisheries and Wildlife	#alaska, alaska, #ak, murkowski, alaskans, anchorage, photo, ak, weekend, air, fairbanks, office, native
Energy (Misc)	energy, gas, natural, oil, clean, resources, forward, #utpol, #energy, looking, wind, forest
Energy (Mining)	#4jobs, energy, epa, bills, gas, @housecommerce, #energy, passed, regulations, gop, #jobs, #stopthetaxhike
Commercial Banks	hearing, committee, financial, subcommittee, oversight, services, reform, @financialcmte, consumer, cmte, chairman, secretary
Securities and Investment	bipartisan, senate, bill, pass, extension, cut, compromise, house, tax, passed, #4jobs, jobs, gop
Credit Unions	tax, ur, mortgage, recession, generation, honest, blog, people, rate, terrorist, don, self
Health Professionals	health, medicare, care, obamacare, #obamacare, reform, republicans, repeal, seniors, healthcare, americans, democrats
Casinos and Gambling	inouye, senator, lujn, hawaii, nevada, heck, joe, nv, berkley, meeting, attending, #hawaii
Pro-Israel	iran, women, rights, nuclear, israel, ben, violence, gop, senate, security, #vawa, cardin
Women's Issues	hagan, nc, women, stabenow, mo, #hawaii, contracting, vets, #mo, #women, game, #nc

Table 2: MIPT's word-industry associations, for some manually selected industries.

Results, averaged across folds, are shown in Table 1. Only SIPT improves over the baseline (statistical significance at $p < 0.001$, Wilcoxon signed rank test, for all metrics). This increased predictability indicates a connection between contribution profiles and public messages. Of course, a causal relationship cannot be inferred (in either direction).

The dramatic difference in predictive performance across models suggests the importance of careful model design. The discriminative model posits a similar word-industry association to our model but ignores the message level, assuming all messages are equally explained proportional to $\boldsymbol{\pi}_c$. MIPT posits a very high dimensional latent structure that may not be learnable from the amount of training data available here. SIPT strikes a better balance.

We found the MIPT model gives *qualitatively* better word-industry associations with greater face validity, despite its inadequacy as a predictor. This is not uncommon in unsupervised topic modeling; similar observations have been made before (Boyd-Graber et al. 2009).

[1] Additional rounds were not helpful in preliminary trials.

[2] This follows from the "d-separation" property observable in Fig. 2: there is no active path between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

[3] The regularized maximum likelihood estimator is: $\arg\max_{\boldsymbol{\theta}} \sum_c \sum_i \pi_{c,i} \log p_{\boldsymbol{\theta}}(i|\boldsymbol{w}_c) + \lambda\|\boldsymbol{\theta}\|_2^2$ where the multinomial logit distribution $p_{\boldsymbol{\theta}}$ is based on the same unigram features considered by our generative models. λ is tuned using the 10% of the training data as the development set.

Table 2 shows words MIPT associates with some industries. We emphasize that these associations were revealed using only campaign contribution data coupled with tweets by MCs. Many terms appear that are topically related to issues of interest to these industries. We also see states where these industries do business (NC/tobacco, AK/fishing), and the names of MCs who head committees relevant to the industry's interests (Harkin/agriculture, Inouye/casinos). Deviations are also interesting; Sen. Hagan's name associates with women's issues (EMILY's List is one of her top donors), but not tobacco, despite her NC constituency. In the energy sector, jobs appear to be of special importance to miners. Subjectively, we found the industry-word associations discovered by SIPT and the discriminative model to be far less interpretable. So while MIPT does not perform well as a predictive model, it more successfully infers human-interpretable associations. We also ran LDA on just the text (without campaign contribution data); the topics were difficult to distinguish from each other.

Related Work

Tweets have gained interest as observational data relevant to social science questions, including demographic associations with linguistic variation (Eisenstein, Smith, and Xing 2011) and the formation of social norms (Kooti et al. 2012). Golbeck et al. (2010) conducted a comprehensive analysis of the use of Twitter by MCs, and White and Counts (2012) incorporated Twitter data in a spatial model of political ideology. Slightly farther afield, Twitter data has also served as a social "sensor" for detecting flu epidemics (Paul and Dredze 2010) and earthquakes (Sakaki, Okazaki, and Matsuo 2010), for measuring public opinion (O'Connor et al. 2010), and (perhaps less successfully) for predicting elections (Tumasjan et al. 2010; Gayo-Avello 2012; Jungherr, Jürgens, and Schoen 2012). The general use of text as data in political science is an active area of research, often making use of latent-variable models like ours (Quinn et al. 2006; Grimmer 2010; Grimmer, Messing, and Westwood 2012). Evaluation of such models through *prediction* was explored in predicting Congressional roll call votes (Gerrish and Blei 2011) and bill survival in Congressional committees (Yano, Smith, and Wilkerson 2012).

Conclusion

Using a probabilistic model, we have found evidence of a connection between what members of Congress say on Twitter and contributions to their campaigns; using another, we can explore the nature of that connection. In future work, we might consider modeling temporal dynamics in contributions and messages, which might give evidence for causal relationships. Social and individual attributes (e.g., region, party, followers) might help identify additional factors that help explain politicians' message content.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet allocation. *JMLR* 3:993–1022.

Boyd-Graber, J.; Chang, J.; Gerrish, S.; Wang, C.; and Blei, D. 2009. Reading tea leaves: How humans interpret topic models. In *NIPS*.

Eisenstein, J.; Smith, N. A.; and Xing, E. P. 2011. Discovering sociolinguistic associations with structured sparsity. In *ACL*.

Gayo-Avello, D. 2012. No, you cannot predict elections with twitter. *IEEE Internet Computing* 16(6):91–94.

Gerrish, S., and Blei, D. 2011. Predicting legislative roll calls from text. In *Proc. of ICML*.

Golbeck, J.; Grimes, J.; and Rogers, A. 2010. Twitter use by the U.S. congress. *Journal of the American Society for Information Science and Technology* 61(8).

Grimmer, J.; Messing, S.; and Westwood, S. 2012. How words and money cultivate a personal vote: The effect of legislator credit claiming on constituent credit allocation. *American Political Science Review* 106(4).

Grimmer, J. 2010. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis* 18(1).

Jungherr, A.; Jürgens, P.; and Schoen, H. 2012. Why the pirate party won the german election of 2009 or the trouble with predictions: A response to Tumasjan et al. *Social Science Computer Review* 30(2):229–234.

Kooti, F.; Yang, H.; Cha, M.; Gummadi, P. K.; and Mason, W. A. 2012. The emergence of conventions in online social networks. In *ICWSM*.

Lerman, K., and Ghosh, R. 2010. Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. In *ICWSM*.

Minka, T. 2000. Estimating a Dirichlet distribution. http://bit.ly/XTEJFu.

Nigam, K.; McCallum, A. K.; Thrun, S.; and Mitchell, T. 1999. Text classification from labeled and unlabeled documents using EM. *Machine Learning* 39:103–134.

O'Connor, B.; Balasubramanyan, R.; Routledge, B. R.; and Smith, N. A. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *ICWSM*.

Paul, M., and Dredze, M. 2010. You are what you tweet : Analyzing Twitter for public health. In *ICWSM*.

Quinn, K. M.; Monroe, B. L.; Colaresi, M.; Crespin, M. H.; and Radev, D. R. 2006. An automated method of topic-coding legislative speech over time with application to the 105th–108th U.S. Senate. Midwest Political Science Association Meeting.

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW*.

Tumasjan, A.; Sprenger, T. O.; Sandner, P. G.; and Welpe, I. M. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *ICWSM*.

Wainwright, M. J., and Jordan, M. I. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* 1(1–2):1–305.

White, J. M., and Counts, S. 2012. Improving spatial models of political ideology by incorporating social network data. In *Workshop on Information in Networks*.

Yano, T.; Smith, N. A.; and Wilkerson, J. D. 2012. Textual predictors of bill survival in congressional committees. In *NAACL*.