Automatic Discovery of Partisan campaigns, Agenda Setting and Political Spin in Press releases

Oren Tsur[†§] orentsur@seas.harvard.edu
Dan Calacci[†] dcalacci@ccs.neu.edu
David Lazer[†‡] d.lazer@neu.edu
[†]Lazer Laboratory, Northeastern University
[§]School of Engeneering and Applied Sciences, Harvard University
[‡]Harvard Kennedy School, Harvard University

ABSTRACT

Press releases (PR) serve as an important tool by which political figures (and business) communicate their messages to news editors and journalists, which in turn deliver it to their audience. Indeed, much of the media coverage of political events is based on, responds to or quotes press releases submitted by politicians. With the increasing number of PR communications it is important to provide journalists with tools for easy processing of press releases at large scale. In this paper we present a system for automatic discovery of agenda setting efforts, framing strategies and political spin as evident in a large corpus of press releases. The system combines topic models, sentiment analysis and autoregressive-distributed-lag models. We automatically analyze over 130000 PR communications released by members of the House and the Senate in the years 2010-2013 and find significant differences in topic ownership and sentiment as well as significant evidence for coordinated campaigns for agenda setting and political spin. We provide a detailed analysis of agenda setting campaigns related to two important issues in American politics: the health care reform and energy related issues.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Computational Journalism, Application, Data Analysis

Keywords

press releases, autoregressive models, distributed-lag models, attention, spin, political science, topic models, sentiment analysis

1. INTRODUCTION

Press Releases (PR) serve as an important mean for effective sharing of relevant information with news outlets. However, editors, journalists and reporters may be overwhelmed by the growing number of press releases they receive daily. Studies show that much of the media coverage of political events is based on, responds to and frequently quotes (or copies) press releases [11, 1]. Advances in computational linguistics and data analysis may be employed in order to enable easier processing of this load of information. Using such tools will result with the media ability to provide their readership/viewers with a more accurate and interesting story as well as increasing transparency and facilitating a better educated constituency.

Language is one of the main tools used by politicians to promote their agenda, gain popularity, win elections and drive societal change [16]. Theoretical constructs employed by Communication and Political Science scholars to describe features of the political communication mechanism include: *topic ownership*, *framing*, and *agenda setting*.

We say that a candidate, a representative or a party *owns* a topic if this topic or set of ideas are associated with her/the party as opposed to their competition. For example, environmental issues are traditionally associated with specific parties (in the case of U.S. politics, environmental issues are mostly associated with the Democratic party).

Politicians can use different *frames* when referring to a specific topic, giving the public very different views on a topic at hand. A notable example is the divisive partisan rhetoric used by U.S. politicians when referring to the highly controversial debate regarding the legality of abortion. Democratic and Republican positions are framed as 'pro choice' and 'pro life', respectively. The different farming spin the issue of abortion as the manifestation of the values of individual freedom or of sanctity of life. Similarly, Republicans frame the U.S. inheritance tax as the overwhelmingly negative 'death tax', while Democrats use the neutral 'estate tax'. The Affordable Care Act (ACA) is often referred to as 'Obamacare' by Republicans framing it as an obsession of a too liberal President[1], while Democratic rhetoric tends to focus on Medicare/Medicaid.

Agenda setting is achieved by framing and increased or decreased *attention* (attention shifts) in order to set or change the political, media or public agenda. An example of agenda

[1]Recently, however, Democrats started using the term 'Obamacare', in attempt to reclaim and re-frame it.

setting may be a political leader raising tension with neighboring countries in order to shift public attention from domestic issues, such as a failing economy.

In the context of this work we refer to the two mechanisms described above (framing and attention shifts) as *political spin*[2]. While we focus on the latter, we are addressing framing indirectly via fine granularity of topic ownership.

In this work we propose a novel framework for unsupervised analysis of public statements issued in the political sphere. We use topic models in order to detect the topics and sub-topics involved in the political discourse. We then determine partisan ownership of topics and the corresponding sentiment. We use autoregressive distributed-lag regression models in order to find significant correlation between topical and sentimental time series. These correlations are then interpreted as evidence for attention shifts and political spin. In the main part of this paper, for illustrative purposes, we focus on seven specific topics that fall under two topical clusters: Health and Energy.

2. RELATED WORK

The success of framing strategies is studied by the analysis of real time reactions to political debates [3]. Expressed agendas in press releases issued by U.S. Senators have been modeled using author topic models [7]. Topic models have been used to detect connections between contributions and political agendas as expressed in microblogging platforms [18] and for reconstructing voting patterns based on the language in congressional bills [6]. Influence and the flow of policy ideas has been modeled via measuring text reuse in different versions of bill proposals [10]. Slant in news articles has been modeled by [14] and [5], comparing word tokens and n-grams to predefined lists and tokens found in labeled data. Hidden Markov Models are used by [17] in order to measure ideological proportions in political speech, and [9] use recursive neural networks for a similar task. Autoregressive models are used for analyzing adjustment of issue positions based on network strategies [12]. Logistic regression on manually coded campaign advertisements is used in order to learn the dynamics of issue ownership by individual candidates [4].

3. DATA

We use a collection of press releases and statements made by U.S. Representatives in the years 2010-2013. These public statements are available via the VoteSmart project[3]. In total we obtained 134000 statements made by 641 representatives (both House and Senate) in a four year span. This time span encompasses two Congressional elections (January 2011 and January 2013) and one Presidential election (November 2012).

4. COMPUTAITONAL FRAMEWORK

In order to automatically discover the correlated dynamics of attention shifts, we take a layered approach, consisting of the following stages:

Stage 1: Topic inference In the first stage, we use topic models in order to learn topic distribution over words and identify the set of topics addressed in the corpus. In practice, we assume a Dirichlet prior on topics and use variational Bayes (VB) optimization to infer topic distributions over words [2]. In order to considerably improve efficiency, we use an online variational Bayes inference algorithm, shown to perform similarly to batch LDA [8].

Stage 2: Data slicing The corpus is sliced according to various parameters in different granularities, and topic distributions are computed by time frame, authorship and partisanship.

Stage 3: Sentiment analysis Once the corpus is sliced, we can assign sentiment scores to statements and topics in various time frames. In this stage, we generate additional time series that describe partisan sentiment towards each topic discovered in stage 1. We use sentiment categories from the LIWC corpus [19].To gain more accurate sentiment results, we adjust sentence-level scores according to sentence structure and negation patterns and normalize scores by document length.

Stage : Autoregressive-Distributed-Lag Models In the last stage we fit a set of regression models to our data in order to estimate the dependency between the time series' obtained in stages 2 and 3. We argue that a lagged dependency between two time series suggests a deliberate attention shift.

We aim to minimize the MSE in the following model:

$$y_t = \alpha + \beta' X_{t'} + \gamma^T Y^n + W^T I(\cdot) + \epsilon_t \quad (1)$$

Where $\beta^T X^m$ indicates the distributed-lag terms, $\gamma^T \cdot Y^n$ indicates the autoregressive component, $t' = t - i$ and $I(\cdot)$ is the identity matrix $I_{12\times 12}$, accounting for seasonality effects ($I_{l,l} = 1$ iff t is in month l). the importance of controlling for autoregressive properties and for seasonality effects was recently demonstrated in the error analysis of the Google Flu Trends algorithm [13]. Following [15] we assume that the news cycle spans up to two weeks, and that attempts of changing the focus of attention are made within that time. To this end, $i \in 0, 1, 2$ indicates no lag, one week lag and 2 weeks lag, respectively.

5. EXPERIMENTS AND RESULTS

5.1 Experimental setting

For the topic model component we use $k = 100$ and allow a document to have 5 topics (ignoring the long tail of statistical noise). Discovered topics were annotated by two annotators based on the top 20 words in each topic. We allowed heirarchial labels (e.g. 'energy' vs. 'energy → cleantech'. Annotator achieved full agreement after consulidation (e.g. 'energy → cleantech' and 'energy → green' are merged). Table 3 contains examples of top words and labels for seven topics in two topical clusters.

We checked topic ownership and partisan sentiment in three different granularities: on a weekly base, a monthly base and over the span of the whole four year.

[2]Fact-twisting and rumor spreading are other forms of political spin. We do not address these in this work.
[3]http://votesmart.org/

For time series analysis we focused on two topical clusters: *Health* and *Energy* (see Table 3). In line with the duration of news cycle reported by [15] we sliced the data on a weekly base. We used the autoregressive-distributed-lag model (described in Stage 4 in Section 4) to find significant correlations between partisan topic ownership and sentiment. For *each* topic we define six *types* time series: weekly normalized counts for Democrats/Republicans and weekly positive/negative sentiment score by Democrats/Republicans. We allow lags of $l \in \{0, 1, 2\}$ weeks.

5.2 Results

Partisan Topic Ownership and Sentiment. Tables 1 and 2 portray partisan topic ownership by the total number of statements and by the number of weeks of ownership, respectively. The tables also illustrate the difference between ownership of a topical cluster and ownership of specific topics within clusters. For example, while Democrats own the health cluster, a closer look at specific topics in the cluster (all automatically inferred by LDA) reveals a more complex picture. Topic 30, the dominant topic in the Health cluster, is actually owned by Republicans both in terms of the total number of statements and in the number of weeks of ownership. Examining the Energy cluster closely reveals the importance of distinguishing between aggregate ownership and weekly ownership. While Republicans issued 514 more statements attributed to the Energy cluster than Democrats (Table 1), Democrats maintained weekly ownership, releasing more statements than Republicans in 107 of the 202 weeks that Energy-related statements were released (Table 2). Similar differences appear when the Energy cluster is split into specific topics.

We also test the significance of ownership differences and deviation of sentiment. We consider a few settings: monthly vs. weekly and cluster level vs. specific topics. The results of the significance tests are summarized in Table 4, most categories show significance with $p < 0.001$.

One interesting observation (Table 4) is that negative sentiment in all settings is dominated by the Republican party. Testing whether the negativity trend characterizes the Republican party in general, is specific to the current political constellation (Obama, Tea Party) or simply characterize the party that is not in office is beyond the scope of this paper.

Other notable results (Table 4) are that ownership loses significance on the cluster level, and that sentiment (both positive and negative) is more significant in the weekly analysis than in the monthly analysis. Aggregated (cluster-level) analyses results in lower significance levels than more granular analyses.

Attention shift, Agenda setting and spinning. As described in Section 5.1, we have thousands of pairs of time series. We fit an autoregressive distributed-lag regression model to each of the pairs and compute the significance of the correlation between the time series. Coefficients of fifty three of these pairs were found to be statistically significant ($p < 0.05$).

Table 5 provides the number of significant correlations of a party (*responsive party*) responding (with lags) to the other party (implicit in the table). Responsiveness is interpreted as correlated reaction to change in attention (*counts*) or to change in sentiment (*sent*). Results are also divided to responses within topical clusters and responses between topical clusters. For example, the bottom right-most entry in the table (marked by a dagger †) indicates three time series in which the Republican party is "responding" to the a change of sentiment (*sent*) of the Democratic Party. The responses appear after a two week lag and are manifested as a shift in attention to an unrelated topic (*inter cluster*). Similarly, the entry marked with ‡ indicates two cases in which the Democratic party responds to an attention shift (*count*) by the Republican party. The responses are lagged by one week. The responses may be attempts to change the framing of a topical cluster(*intra topic*).

Cluster	Topic	DEM	REP	DEM	REP
Health	30	7228	9884*	14280*	13279
	51	2424*	1283		
	80	4628*	2112		
Energy	38	348	861*	7801	8315*
	47	237*	88		
	69	2271	2326*		
	71	4945	5040*		

Table 1: Total number of statements by party in two topical clusters. DEM indicates the Democrat party, REP indicates the Republican party.

Cluster	Topic	DEM	REP	DEM	REP
Health	30	62	139*	125*	82
	51	155*	39		
	80	182*	19		
Energy	38	50	87*	107*	95
	47	88*	25		
	69	105*	84		
	71	107*	91		

Table 2: Number of *weeks* each party "owned" a topic.

It appears that the Democratic party is more responsive to both attention shifts and changes in sentiment within and between clusters. We attribute the Democratic responsiveness to the fact that the Republican party holds majority in the House during most of the span of the data.

Twenty five of the correlated time series are found for $lag = 0$. No-lag correlations suggest that there is a responsive pattern, but due to the granularity of our analysis, it is impossible to determine which party is 'setting' the agenda, and which party is practicing agenda 'shifting'[4].

6. CONCLUSION

In this paper we proposed a novel framework for unsupervised analysis of press releases and statements issued in the political sphere. We employ topic models in order to detect

[4]In the cases of $lag = 0$, the interpretation of the "Responsive Party" is by controlling for autoregressive error of this party's time series.

Cluster	Topic ID	Top Words	Labels
Health	30	health care law will obamacare insurance repeal affordable americans costs new reform people president healthcare act coverage mandate american obama	healthcare reform
	51	medicare seniors program social medicaid benefits fraud payments security programs cost services costs billion payment beneficiaries waste year savings million	health budget
	80	health care insurance medical drug will coverage drugs patients prescription new fda access hospitals act hospital affordable companies help can	healthcare reform
Energy	38	project pipeline president obama keystone jobs climate energy xl construction state change permit administration approval oil will canada environmental create	energy economy keystone-xl
	47	companies big corporations corporate oil profits romney loopholes overseas breaks interests ship bachmann subsidies profit loophole buy close candidate mitt	energy economy politics
	69	oil alaska gulf coast spill drilling offshore bp murkowski begich markey resources noaa said industry moratorium mexico gas administration sen	energy environment
	71	energy gas oil production will prices clean american power natural fuel jobs nuclear domestic new resources renewable wind development america	energy clean-tech

Table 3: Top twenty words and hierarchical labels for the topics in the Health and the Energy clusters.

Cluster	Topic	Own_w	Own_m	$+_w$	$-_w$	$+_m$	$-_m$	Own_w	Own_m	$+_w$	$-_w$	$+_m$	$-_m$
Health	30	R^{*}	—	D^{***}	R^{***}	D^{***}	R^{***}						
	51	D^{***}	D^{***}	D^{***}	—	D^{***}	—	—	—	D^{***}	R^{***}	D^{***}	R^{***}
	80	D^{***}	D^{***}	D^{***}	R^{***}	D^{***}	R^{***}						
Energy	38	R^{*}	R^{*}	R^{*}	—	R^{**}	—						
	47	D^{***}	D^{**}	—	—	—	—	$D^{\dagger}$	—	D^{*}	R^{***}	—	R^{***}
	69	—	—	R^{**}	—	R^{***}	—						
	71	—	—	D^{**}	R^{***}	—	R^{*}						

Table 4: Topic ownership (Own) and sentiment (positive $+$, and negative $-$) by topic, topical cluster and party. Weekly and monthly aggregates are indicated by w (weekly) and m (monthly) in $< Own| + |- >_{w,m}$ (Weekly columns are grayed out). Asterisks indicate significance: (***) indicates $p < 0.001$, (**) indicates $p < 0.01$ and (*) indicates $p < 0.05$. Empty cells indicate statistically insignificant ($p > .1$) results using an independent two-sample t-test. A dagger (†) indicates $.05 < p < .1$.

Cluster	Responsive Party	Independent Variable (Party)					
		Count			Sent		
		$lag = 0$	$lag = 1$	$lag = 2$	$lag = 0$	$lag = 1$	$lag = 2$
Intra-cluster	D	4	$2^{\ddagger}$	0	3	3	0
	R	4	1	0	0	2	3
Inter-cluster	D	4	3	3	2	5	1
	R	4	0	1	4	1	$3^{\dagger}$

Table 5: Number of statistically significant ($p < .05$) correlations between cross-party time series. Topics within a single topical cluster are 'Intra-Cluster' correlations, while those between topic clusters are 'Inter-Cluster' correlations. 'Responsive Party' refers to the party whose statement counts/sentiment we use as our dependent variable in the regression. The responsive party is always responding to the other party (not indicated explicitly). Results are divided into columns based on whether the response is for a change in attention or for change in sentiment. Results are further divided by the lag term used in that correlation test.

topics expressed in political discourse. We then determine partisan ownership of topics and corresponding sentiment. After sanitizing the data, we find significant correlations between topical and sentimental time series using autoregressive distributed-lag regression models. These correlations are then interpreted as evidence for attention shifts and political spin.

Applying this framework on a set of more than 130000 statements and press releases we automatically find statistically significance sub-topic ownership corresponding to framing efforts, we find a striking difference in partisan sentiment and we find lagged correlations that are interpreted as campaigns for agenda setting and political spin.

Future work includes several lines of research auch as (1) moving beyond the macro level of partisanship to the micro level of individual representatives to detect political influence and idea propagation in the political networks, and (2) incorporating other, less formal, mediums of communication commonly used by representatives, such as interviews (TV, radio, papers) and social media outlets.

7. ACKNOWLEDGMENTS

We thank Katherine Ognyanova, Brian Keegan and Ryan Kennedy for fruitful discussions and their helpful comments. We thank Lisa Friedland for sanitizing the statement categories pulled from the VoteSmart Project. This work was supported by the following grants: MURI #504026 and ARO #50433.

8. REFERENCES

[1] How news happens: A study of the news ecosystem of one american city. *Pew Research Center Report*, 2010.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[3] A. E. Boydstun, R. A. Glazier, M. T. Pietryka, and P. Resnik. Real-time reactions to a 2012 presidential debate a method for understanding which messages matter. *Public Opinion Quarterly*, 78(S1):330–343, 2014.

[4] D. F. Damore. The dynamics of issue ownership in presidential campaigns. *Political Research Quarterly*, 57(3):391–397, 2004.

[5] M. Gentzkow and J. M. Shapiro. What drives media slant? evidence from u.s. daily newspapers. *Econometrica*, 78:35–71, 2010.

[6] S. Gerrish and D. M. Blei. How they vote: Issue-adjusted models of legislative behavior. In *Neural Information Processing Systems (NIPS)*, pages 2762–2770, 2012.

[7] J. Grimmer. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35, 2010.

[8] M. Hoffman, F. R. Bach, and D. M. Blei. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864, 2010.

[9] M. Iyyer, P. Enns, J. Boyd-Graber, and P. Resnik. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd meeting of the Association for Computational Linguistics*, 2014.

[10] D. A. S. John Wilkerson and N. Stramp. Tracing the flow of policy ideas in legislatures: A text reuse approach. In *New Directions in Analyzing Text as Data*, Sept. 2013.

[11] S. Kiousis, M. Mitrook, X. Wu, and T. Seltzer. First-and second-level agenda-building and agenda-setting effects: Exploring the linkages among candidate news releases, media coverage, and public opinion during the 2002 florida gubernatorial election. *Journal of Public Relations Research*, 18(3):265–285, 2006.

[12] J. Kleinnijenhuis and W. de Nooy. Adjustment of issue positions based on network strategies in an election campaign: A two-mode network autoregression model with cross-nested random effects. *Social Networks*, 35(2):168–177, 2013.

[13] D. M. Lazer, R. Kennedy, G. King, and A. Vespignani. The parable of google flu: Traps in big data analysis. *Science Magazine (AAAS)*, 2014.

[14] H. S. Lee. Do national economic and political conditions affect ideological media slant? *Political Communication*, 30:395–418, 2013.

[15] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. ACM, 2009.

[16] F. I. Luntz. *Words That Work: It's Not What You Say, It's What People Hear*. Hyperion, 2007.

[17] Y. Sim, B. Acree, J. H. Gross, and N. A. Smith. Measuring ideological proportions in political speeches. In *Proceedings of EMNLP*, 2013.

[18] D. Y. Tae Yano and N. A. Smith. A penny for your tweets: Campaign contributions and capitol hill microblogs. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, July 2013.

[19] Y. R. Tausczik and J. W. Pennebaker. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54, Mar. 2010.