Predicting Government (Non)Responsiveness to Freedom of Information Requests with Supervised Latent Dirichlet Allocation

Benjamin E. Bagozzi
Department of Political Science
& International Relations
University of Delaware
bagozzib@udel.edu

Daniel Berliner
School of Politics
& Global Studies
Arizona State University
danberliner@gmail.com

Zack W. Almquist
Department of Sociology
& School of Statistics
University of Minnesota
almquist@umn.edu

Abstract

Understanding government responsiveness to citizen information requests is important to theories of political accountability, as well as to practitioners' abilities to monitor and improve this crucial transparency mechanism. We use supervised latent Dirichlet allocation techniques to predict the Mexican government's (non)responsiveness to *all* federal information requests filed during the period 2003-2015. After presenting our approach, we assess its value-added in both the in-sample and out-of-sample settings.

1 Introduction

Following Mexico's landmark 2002 access to information law (Berliner and Erlich, 2015), every single freedom of information request filed with federal government agencies has been made publicly available—now over one million requests in total. Understanding the Mexican government's responsiveness to these individual information requests is important for theories of government responsiveness (and its politicization), as well as for practitioners' abilities to monitor, scrutinize, and improve the quality of this critical accountability mechanism. Such laws, similar to the Freedom of Information Act in the United States, have now been adopted by over 100 countries around the world (Berliner, 2014; Berliner, 2016).

After converting the complete corpus of Mexican public information requests (2003-2015) into machine readable text, we use topic models to predict government (non)responsiveness towards Mexican information requests in both an in-sample and out-of-sample context. Specifically, we apply supervised latent Dirichlet allocation (sLDA) techniques to this text corpus, so as to evaluate the extent to which one can use the texts of individual requests to predict the (i) time until a government response and (ii) probability of a "denied request." We then evaluate the value-added of this approach against several alternatives. Finally, we assess our sLDA topics for their "politicization," and find that the topics that are most strongly associated with *nonresponsiveness* do indeed exhibit more politicization than do the topics most associated with high *responsiveness*.

2 Background

Democratic institutions are founded on the notion of responsive government, but responsiveness is usually limited and incomplete. Many scholars have studied why political actors may be more or less responsive in different circumstances — both at a macro-scale in terms of how government policies and spending respond to the preferences of the median voter (Golden and Min, 2013), and at a micro-scale in terms of individual citizen-government interactions (Lagunes, 2008; Butler and Broockman, 2011; McClendon, 2016).

Building upon the latter approach, we examine government responsiveness in one case of frequent government-citizen interaction: responses to public information requests in Mexico. To do so, we use a comprehensive dataset of over one million information requests filed with federal government agencies. These correspond to queries made by individual citizens, legal representatives, businesses, and NGOs to specific Mexican federal government agencies, and cover, for example, requests for information on government salaries, land use and zoning restrictions, or distributive programs. Due to the unique online information platform created by Mexico's 2002 *Ley Federal de Transparencia y Acceso a la Información Pública Gubernamental*, the text of each of these requests, along with associated metadeta, has been made publicly available for the years 2003-2015.

Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText 2016), pages 8–13, Dubrovnik, Croatia, 14–16 July 2016.

2.1 Measuring (Non)Responsiveness

Our analysis focuses upon predicting government (non)responsiveness to these information requests. We are interested both in the *timing* of response and in the *type* of response: information provided or denied. We accordingly use two separate measures to evaluate the (non)responsiveness to any given information request: (i) a binary indicator of "denied requests" (for various reasons) and (ii) information on the time-until-response.

Regarding our time-until-response measure, we create an outcome variable that corresponds to the logged number of working days (excluding weekends and official Mexican government holidays) until an information request response is provided to the requestor by the Mexican government. While the standard time limit for the Mexican government to provide a response is 20 working days, officials can request an extension of up to a maximum of 40 working days. Across our entire dataset, 66.4% of requests received responses within 20 working days while 89.3% of requests received responses within 40 working days. Our final (logged) time-until-response measure has a mean of 2.89 and range of 0.00-to-7.59.

Our binary "denied request" indicator is our primary outcome of interest, and is based upon the coding scheme developed by Fox et al. (2011), which classifies any response marked as "No es de competencia," "Inexistencia," "Reservada," "No se dará trámite," "Solicitud no corresponde al marco de la ley" and "Sin Respuesta" as a "denied request" ($= 1$), and zero otherwise. The resultant "denied request" indicator is moderately imbalanced with a sample mean of 0.23. Finally, we then also omit the final two months of information requests from our analyses below, to ensure that we do not treat any cases marked as "Sin Respuesta" as "denied" when they had simply not yet exceeded the time limits for response.

2.2 Information Request Features

We focus on the request texts themselves as our primary features of interest. These texts correspond to each requestor's own open-ended description of the specific information that they are requesting. Because public officials are the primary responders to these requests, we believe that the themes found across these requests, and their varying degrees of politicization, will help to predict government (non)responsiveness.

We thus downloaded all requests from Mexico's online information request interface. While most requestors described the nature of their requests within the designated field, a smaller subset (roughly 13%) included a portion or all of their request as an attachment. Because these attachments are relevant to our analysis, we additionally downloaded each attachment and added these into our primary request text field, along with any auxiliary request content. We then (i) removed all requests pertaining to confidential personal information and (ii) truncated all remaining requests from the thousandth string onwards.[1] This created our primary corpus of interest, which was further preprocessed using standard approaches (e.g., stemming) for the automated analysis of political texts (Bagozzi and Schrodt, 2012; Bagozzi, 2015; Berliner et al., 2016). Altogether, the above steps yielded a corpus of 1,003,756 requests.

We next appended the names of each request's designated federal government agency to our processed texts. Each information request in our sample designated a single government agency, such as the Instituto-Nacional-de-Desarrollo-Social, as the *target agency* for the information that was requested. As these agencies vary in their levels of politicization and resources, we anticipate agency-designation, like a request's textual content, to influence the degree of (non)responsiveness to a given request. Agency information was included as an additional field within the original request metadata, and encompasses roughly 300 distinct Mexican federal agencies for our sample. Further below, we evaluate the contribution of this addition to our prediction and classification tasks.

3 Supervised Latent Dirichlet Allocation

Topic models have recently been shown to be highly valid for the discovery of latent thematic content within Mexico's information request texts (Berliner et al., 2016). As such, the present paper evaluates the utility of *supervised* latent Dirichlet allocation (sLDA) models (Blei and Mcauliffe, 2008) for the prediction of government (non)responsiveness to these same request texts.

sLDA is a probabilistic topic model designed for identifying the groupings of words that are most predictive of a document-indexed response variable. sLDA estimates these groupings of

[1] Only 0.02% of our documents have more than 1,000 strings; most are attachments with extensive itemized lists.

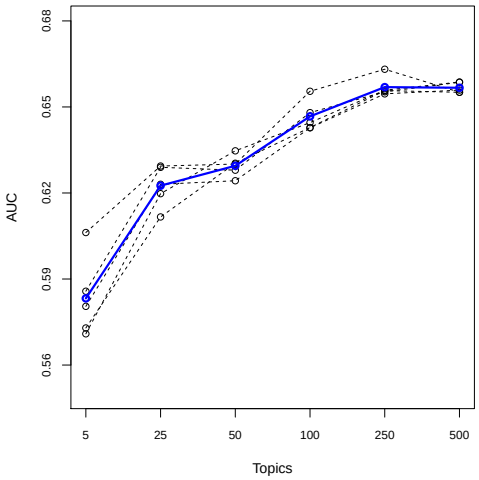
words—hereafter referred to as topics—via a three-level hierarchical model that treats each document as containing a finite mixture of underlying topics, where the topics themselves are specified as an infinite mixture over a corresponding latent set of topic probabilities. One's document-level responses are then regressed on these estimated topic frequencies so as to restrict responses to be non-exchangeable with words, while allowing for flexibility between topic frequency and response type under a generalized linear model (GLM) framework (Blei and Mcauliffe, 2008).

Under this approach, our information request texts are assumed to be mixtures of multiple *latent topics*, each with a characteristic set of words. We anticipate that a subset of these latent topics will be highly politicized, and hence expect that our modeling of all topics across all request documents will aid in the prediction of government (non)responsiveness, as measured via (i) logged time-until-response or (ii) "denied request." In each sLDA model presented below, we specify the distribution of the former response variable to be Gaussian and the latter to be logistic, and perform estimation using collapsed Gibbs sampling via the 'lda' package in R (Chang, 2015).

Researchers must assign the number of topics, k, to be estimated within sLDA. We use a five-fold cross-validation approach to identify an optimal number of topics for the task of prediction. To do so, we first draw a random sample of approximately 250,000 information requests and then randomly partition this sample into five folds of training and test data. For each set of training data, we next estimate a series of sLDA models where the number of topics, k, is sequentially set to $k = \{5, 20, 50, 100, 250, 500\}$ and where our outcome variable is assigned as the binary "denied request" measure described above. We then use each resultant sLDA model's output to initialize a validation sLDA model using each fold's corresponding test sample. With these results in hand, we calculate the area under each test sample's corresponding receiver operating characteristic curve (i.e., the AUC) for "denied requests."

Figure 1 plots the corresponding AUCs for all k's evaluated, along with mean AUCs (the solid line), and indicates that an optimal number of topics for the task of predicting "denied requests" rests somewhere in the $k = 250$ range, since this topic number yields the highest average AUC for our cross-validation sample (i.e., 66%). We hence set $k = 250$ for all primary sLDA models below.

Figure 1: Cross-Validation Results



4 Evaluations

Having used a random sample of 25% of all request texts to identify an optimal number of topics, we next evaluate our sLDA model on our remaining (held out) texts. To do so, we first re-estimate a final ($k = 250$) sLDA model on *all* of the previously sampled 250,000 documents, separately for each outcome of interest: (i) logged time-until-response and (ii) "denied requests." We then generate in-sample and out-of-sample predictions for our two outcome variables, where for our out-of-sample predictions we use the remaining 75% of our sample data (i.e., $\approx 750,000$ request texts).

4.1 In-Sample Results

In order to assess our in-sample sLDA results for both (i) time-until response and (ii) "denied request," this subsection first discusses our topic-specific coefficient estimates, followed by an evaluation of the topics most predictive of (non)responsiveness, and then finally an assessment of in-sample classification. For both models, nearly all of our 250 topic-specific estimates are statistically significant under traditional thresholds, with the vast majority implying either an increase in responsiveness—or a slight increase nonresponsiveness—when present. However, a small number of topics exhibit very large positive effects on non-responsiveness in each model. We hence identify the two topics with the largest es-

timated effects on (i) nonresponsiveness and (ii) responsiveness from *each* sLDA response-model for further examination.

The top words associated with these 'highly predictive topics' are presented below, where we have de-stemmed all topwords, removed *target agency* names (if present), and translated each resultant word to English. The two topics that are most predictive of nonresponsiveness, Slowest$_{\#1}$ and Denied$_{\#1}$, each capture the same highly politicized theme: investigative requests pertaining to financial improprieties, accreditations, and scandals (e.g., FICREA, a collapsed credit union under fraud investigation). Notably, our sLDA estimates imply that requests associated with this topic see a 18,064% increase in the odds of a " denied request," and a 47 day increase in time-until-response. By comparison, the median increase in the odds of a "denied request," and the median increase in time-until-response—across all 250 of our topic estimates—are 110.7% and 1-day.

The second most predictive topic of a "denied request" (Denied$_{\#2}$), likewise appears to be highly politicized, with topwords associated with inquires into money-and-politics, including topwords such as "money," "where," "diputados" and "senators," and with an estimated increase in the odds of a "denied request" of 13,466%. By contrast, Slowest$_{\#2}$ instead appears to be slightly less politicized with its topwords suggesting a more general focus on government accreditation and endorsement. Nevertheless, on the whole, these four topics are far more politicized than the topwords found within Fastest$_{\#1}$, Fastest$_{\#2}$, Provided$_{\#1}$, Provided$_{\#2}$, which as can be seen below, encompass themes of politeness, benign information queries, and requests concerning commercial-product and energy-rate information.

Topics most predictive of time-until-response:

- Slowest$_{\#1}$: saving, FICREA, financial, users, CONDUSEF, bank, settlement, value, accreditation, society
- Slowest$_{\#2}$: documents, accreditation, published, any, electronic, endorses, I request, copy, contains, fact
- Fastest$_{\#1}$: do, requirements, business, can, answer, necessary, respect, you can, question, information
- Fastest$_{\#2}$: registry, brand, involved, find, commercial, I request, property, kind, medium, so

Topics most predictive of a "denied request":

- Denied$_{\#1}$: value, saving, settlement, financial, protections, any, interventions, concept, banking, society
- Denied$_{\#2}$: money, change, deputies, decommissioned, quantity, western, where, year, information, senators
- Provided$_{\#1}$: electronic, energy, CFE, municipality, consumption, rate, lighting, bills, users, latest
- Provided$_{\#2}$: IFAI, I request, information, published, cape, carry, process, following, opinion, federal

We next evaluate the in-sample classification performance of our sLDA models. In the interest of space, we focus all ensuing discussions on the binary "denied request" outcome and results. We then construct two random "coin-flip" baselines for comparison, hereafter denoted ξ, with the first generating random binary data with probability $\frac{1}{2}$, and the second generating random binary data with probability equal to the mean of our true binary response $\bar{y} = 0.23$. In this manner $\xi = \bar{y}$ provides us with a random classifier that maximizes overall accuracy, whereas $\xi = \frac{1}{2}$ provides us with a random classifier that instead favors the improved identification of cases within our less frequent outcome (i.e., nonresponsiveness).

We compare these two random classifiers against our in-sample "denied request" sLDA results with the aid of AUCs, true positive rates (TPRs), true negative rates (TNRs), F1 scores, and overall classification accuracy. Given our preference for the accurate prediction of our minority class (i.e., nonresponsiveness), we assign a cutoff of 0.25 for the calculation of our TPR, TNR, F1 score, and accuracy values.

As can be seen in Table 1, our AUC values imply that our sLDA in-sample predictions are moderately better than chance (AUC$= 66.49$)—which is a finding that is further reinforced by our sLDA model's superior F1 score and TPR values to those obtained under either $\xi = \frac{1}{2}$ or $\xi = \bar{y}$. As expected, $\xi = \bar{y}$ maximizes overall accuracy, with a value (64.34) that is superior to that of $\xi = \frac{1}{2}$ (50.06). However, the maximized accuracy obtained under $\xi = \bar{y}$ still falls slightly below that of our sLDA classifier (66.10), and comes at the cost of noticeably poorer TPR performance than either $\xi = \frac{1}{2}$ or sLDA, which as mentioned above, is valued more so than TNR in this application given our primary interest in *nonresponsiveness*.

Table 1: In-Sample Classification Statistics

	AUC	**TPR**	**TNR**	**F1score**	**Accuracy**
sLDA	66.49	52.54	70.18	41.77	66.10
$\xi = \frac{1}{2}$	50.04	50.02	50.07	31.67	50.06
$\xi = \bar{y}$	50.04	22.86	76.83	22.87	64.34

4.2 Out-of-Sample Results

We now turn to an evaluation of our sLDA model's out-of-sample classification properties. For this evaluation, we use our primary sLDA model to generate "denied request" predictions for the remaining 75% (i.e., $\approx 750,000$ documents) within our 2003-2015 request sample. Using these predictions, we then repeat the same steps as above in generating two random classifiers for comparison, $\xi = \frac{1}{2}$ and $\xi = \bar{y}$, and recalculate the previously described set of classification statistics, in Table 2.

Table 2: Out-of-Sample Classification Statistics

	AUC	TPR	TNR	F1score	Accuracy
sLDA	66.24	52.26	70.24	41.64	66.08
$\xi = \frac{1}{2}$	50.05	50.06	50.04	31.70	50.04
$\xi = \bar{y}$	50.05	22.95	77.10	23.07	64.56

Our out-of-sample results are largely consistent with our in-sample findings. As above, the sLDA model outperforms both random classifiers in AUC, TPR, F1 score, and overall accuracy, and performs second best (to $\xi = \bar{y}$) in TNR. The results reported in Table 2—across all classifiers—suggest that our out-of-sample sLDA predictions perform comparably to, albeit slightly worse than, our in-sample sLDA results. For example, our sLDA model accurately classifies 66.08% of all out-of-sample cases, whereas in the in-sample context our sLDA model's overall accuracy was 66.10%. Differences between these two sets of sLDA predictions are slightly larger when one examines AUCs (66.49 vs. 66.25), though these differences are again fairly negligible, especially relative to the effect of k on our AUCs in Figure 1.

Finally, though not reported here, we also compared these results to a "requests only" sLDA model that omits our *target agency* names as features, and found that the latter performs slightly worse than our full sLDA model. For example, the "requests only" model's out-of-sample AUC is 64.95, which is noticeably smaller than that of our primary sLDA model. Our remaining comparison metrics yielded similar conclusions: the addition of *target agency* names to our text features leads to a small but consistent improvements in accuracy.

4.3 Comparison to Alternate Approaches

We next compare our sLDA approach to three widely used alternatives: support vector machines (SVMs), logistic regression with Lasso, and random forests (RF). All three of these alternate approaches encountered computational difficulties when applied to our full training set of 250,000 documents, leading us to evaluate each of these classifiers, and sLDA, on a smaller training set ($n = 50,000$) and smaller test set ($n = 150,000$) of documents for the purposes of comparison. The results from this exercise appear in Table 3.

Table 3: Out-of-Sample Comparisons

	AUC	TPR	TNR	F1score	Accuracy
sLDA	65.84	50.28	70.99	40.71	66.21
SVM	65.27	26.54	88.53	32.21	74.23
Lasso	65.39	30.52	86.39	34.70	73.50
RF	70.23	48.52	78.82	44.29	71.83

In Table 3, sLDA performs slightly better than SVM and Lasso—but noticeably worse than RF—in terms of AUC. More generally, SVM and Lasso each appear to under-predict "denied requests," thereby ensuring that these two classifiers have higher TNR and higher overall accuracy than either sLDA or RF, albeit at the expense of worse performances on TPR and F1 score. While RF does exhibit a slightly worse TPR than sLDA, its higher F1 score, higher overall accuracy, and higher AUC suggest that RF outperforms sLDA along most dimensions of comparison, though, on the whole, both approaches (i.e., sLDA and RF) generally outperform Lasso and SVM in Table 3.

5 Conclusion

The content of Mexico's information requests, when modeled with sLDA, can help to predict government (non)responsiveness. Evidence from this exercise further suggests that politicization may increase nonresponsiveness. Future work should refine our approach so as to better accommodate (i) the imbalance in "denied request" outcomes, (ii) additional features (such as a requestor's home municipality), and (iii) the non-hierarchal structure of the Mexican information request data; while also better benchmarking our request text sLDA-classification results against alternative supervised machine learning techniques.

Acknowledgments

Almquist's research was supported in part by ARO YIP Award #W911NF-14-1-0577. He is currently a visiting scholar at the University of Washington.

References

Benjamin E. Bagozzi. 2015. The Multifaceted Nature of Global Climate Change Negotiations. *The Review of International Organizations*, 10(4): 439-464.

Benjamin E. Bagozzi and Philip A. Schrodt. 2012. The Dimensionality of Political News Reports. *Proceedings of the European Political Science Association Meetings*.

Daniel Berliner. 2014. The Political Origins of Transparency *The Journal of Politics*, 76(2): 479-491.

Daniel Berliner. 2016. Transnational advocacy and domestic law: International NGOs and the design of freedom of information laws *Review of International Organizations*, 11(1): 121-144.

Daniel Berliner and Aaron Erlich. 2015. Competing for Transparency: Political Competition and Institutional Reform in Mexican States *American Political Science Review*, 109(1): 110-128.

Daniel Berliner, Benjamin E. Bagozzi, and Brian Palmer-Rubin. 2016. What Information Do Citizens Want?: Evidence from One Million Information Requests in Mexico, Working Paper.

David M. Blei and Jon D. Mcauliffe. 2008. Supervised Topic Models *Advances in Neural Information Processing Systems*, 20:121-128.

Daniel M. Butler and David E. Broockman. 2011. Do Politicians Racially Descriminate Against Constituents? A Field Experiment on State Legislators *American Journal of Political Science*, 55(3):463-477.

Jonathan Chang. 2015. Package 'lda'. *https://cran.r-project.org/web/packages/lda/*.

Jonathan Fox, Libby Haight, and Brian Palmer-Rubin. 2011. Proporcionar transparencia ¿Hasta qué punto responde el gobierno mexicano a las solicitudes de información pública? *Gestión y Política Pública*, 20(1):3-61.

Miriam Golden and Brian Min. 2013. Distributive Politics Around the World *Annual Review of Political Science*, 16(1):73-99.

Gwyneth McClendon. 2016. Race and Responsiveness: A Field Experiment with South African Politicians *The Journal of Experimental Political Science*, 3(2).

Paul Lagunes. 2008. Irregular Transparency? An Experiment Involving Mexico's Freedom of Information Law, Working Paper.