Predicting Response to Political Blog Posts with Topic Models

Tae Yano   William W. Cohen   Noah A. Smith
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
`{taey,wcohen,nasmith}@cs.cmu.edu`

Abstract

In this paper we model discussions in online political blogs. To do this, we extend Latent Dirichlet Allocation (Blei et al., 2003), in various ways to capture different characteristics of the data. Our models jointly describe the generation of the primary documents (posts) as well as the authorship and, optionally, the contents of the blog community's verbal reactions to each post (comments). We evaluate our model on a novel comment prediction task where the models are used to predict which blog users will leave comments on a given post. We also provide a qualitative discussion about what the models discover.

1   Introduction

Web logging (blogging) and its social impact have recently attracted considerable public and scientific interest. One use of blogs is as a community discussion forum, especially for political discussion and debate. Blogging has arguably opened a new channel for huge numbers of people to express their views with unprecedented speed and to unprecedented audiences. Their collective behavior in the blogosphere has already been noted in the American political arena (Adamic and Glance, 2005). In this paper we attempt to deliver a framework useful for analyzing text in blogs quantitatively as well as qualitatively. Better blog text analysis could lead to better automated recommendation, organization, extraction, and retrieval systems, and might facilitate data-driven research in the social sciences.

Apart from the potential social utility of text processing for this domain, we believe blog data is worthy of scientific study in its own right. The spontaneous, reactive, and informal nature of the language in this domain seems to defy conventional analytical approaches in NLP such as supervised text classification (Mullen and Malouf, 2006), yet the data are rich in argumentative, topical, and temporal structure that can perhaps be modeled computationally. We are especially interested in the semi-causal structure of blog discussions, in which a post "spawns" comments (or fails to do so), which meander among topics and asides and show the personality of the participants and the community.

Our approach is to develop probabilistic models for the generation of blog posts and comments jointly within a blog site. The model is an extension of Latent Dirichlet Allocation (Blei et al., 2003). Unsupervised topic models can be applied to collections of unannotated documents, requiring very little corpus engineering. They can be easily adapted to new problems by altering the graphical model, then applying standard probabilistic inference algorithms. Different models can be compared to explore the ramifications of different hypotheses about the data. For example, we will explore whether the contents of posts a user has commented on in the past and the words she has used can help predict which posts she will respond to in the future.

The paper is organized as follows. In §2 we review prior work on topic modeling for document collections and studies of social media like political blogs. We then provide a qualitative characterization of political blogs, highlighting some of the features we believe a computational model should capture and discuss our new corpus of political blogs (§3). We present several different candidate topic models that aim to capture these ideas in §4. §5 shows our empirical evaluation on a new comment prediction task and a qualitative analysis of the models learned.

2   Related Work

Network analysis, including citation analysis, has been applied to document collections on the Web (Cohn and Hofmann, 2001). Adamic and Glance (2005) applied network analysis to the political bl-

ogosphere. The study modeled the large, complex structure of the political blogosphere as a network of hyperlinks among the blog sites, demonstrated the viability of link structure for information discovery, though their analysis of text content was less extensive. In contrast, the text seems to be of interest to social scientists studying blogs as an artifact of the political process. Although attempts to quantitatively analyze the contents of political texts have been made, results from classical, supervised text classification experiments are mixed (Mullen and Malouf, 2006; Malouf and Mullen, 2007). Also, a consensus on useful, reliable annotation or categorization schemes for political texts, at any level of granularity, has yet to emerge.

Meanwhile, latent topic modeling has become a widely used unsupervised text analysis tool. The basic aim of those models is to discover recurring patterns of "topics" within a text collection. LDA was introduced by Blei et al. (2003) and has been especially popular because it can be understood as a generative model and because it discovers understandable topics in many scenarios (Steyvers and Griffiths, 2007). Its declarative specification makes it easy to extend for new kinds of text collections. The technique has been applied to Web document collections, notably for community discovery in social networks (Zhang et al., 2007), opinion mining in user reviews (Titov and McDonald, 2008), and sentiment discovery in free-text annotations (Branavan et al., 2008). Dredze et al. (2008) applied LDA to a collection of email for summary keyword extraction. The authors evaluated the model with proxy tasks such as recipient prediction. More closely related to the data considered in this work, Lin et al. (2008) applied a variation of LDA to ideological discourse.

A notable trend in the recent research is to augment the models to describe non-textual evidence alongside the document collection. Several such studies are especially relevant to our work. Blei and Jordan (2003) were one of the earliest results in this trend. The concept was developed into more general framework by Blei and McAuliffe (2008). Steyvers et al. (2004) and Rosen-Zvi et al. (2004) first extended LDA to explicitly model the influence of *authorship*, applying the model to a collection of academic papers from CiteSeer. The model combined the ideas from the mixture model proposed by Mc-

Callum (1999) and LDA. In this model, an abstract notion "author" is associated with a distribution over topics. Another approach to the same document collection based on LDA was used for citation network analysis. Erosheva et al. (2004), following Cohn and Hofmann (2001), defined a generative process not only for each word in the text, but also its citation to other documents in the collection, thereby capturing the notion of *relations* between the document into one generative process. Nallapati and Cohen (2008) introduced the Link-PLSA-LDA model, in which the contents of the citing document and the "influences" on the document (its citations to existing literature), as well as the contents of the cited documents, are modeled together. They further applied the Link-PLSA-LDA model to a blog corpus to analyze its cross citation structure via hyperlinks.

In this work, we aim to model the data *within* blog conversations, focusing on comments left by a blog community in response to a blogger's post.

3 Political Blog Data

We discuss next the dataset used in our experiments.

3.1 Corpus

We have collected blog posts and comments from 40 blog sites focusing on American politics during the period November 2007 to October 2008, contemporaneous with the presidential elections. The discussions on these blogs focus on American politics, and many themes appear: the Democratic and Republican candidates, speculation about the results of various state contests, and various aspects of international and (more commonly) domestic politics. The sites were selected to have a variety of political leanings. From this pool we chose five blogs which accumulated a large number of posts during this period: Carpetbagger (CB),[1] Daily Kos (DK),[2] Matthew Yglesias (MY),[3] Red State (RS),[4] and Right Wing News (RWN).[5] CB and MY ceased as independent bloggers in August 2008.[6] Because

	MY	RWN	CB	RS	DK
Time span (from 11/11/07)	–8/2/08	–10/10/08	–8/25/08	–6/26/08	–4/9/08
# training posts	1607	1052	1080	2045	2146
# words (total)	110,788	194,948	183,635	321,699	221,820
(on average per post)	(68)	(185)	(170)	(157)	(103)
# comments	56,507	34,734	34,244	59,687	425,494
(on average per post)	(35)	(33)	(31)	(29)	(198)
(unique commenters, on average)	(24)	(13)	(24)	(14)	(93)
# words in comments (total)	2,287,843	1,073,726	1,411,363	1,675,098	8,359,456
(on average per post)	(1423)	(1020)	(1306)	(819)	(3895)
(on average per comment)	(41)	(31)	(41)	(27)	(20)
Post vocabulary size	6,659	9,707	7,579	12,282	10,179
Comment vocabulary size	33,350	22,024	24,702	25,473	58,591
Size of user pool	7,341	963	5,059	2,789	16,849
# test posts	183	113	121	231	240

Table 1: Details of the blog data used in this paper.

our focus in this paper is on blog posts and their comments, we discard posts on which no one commented within six days. We also remove posts with too few words: specifically, we retain a post only if it has at least five words in the main entry, and at least five words in the comment section. All posts are represented as text only (images, hyperlinks, and other non-text contents are ignored). To standardize the texts, we remove from the text 670 commonly used stop words, non-alphabet symbols including punctuation marks, and strings consisting of only symbols and digits. We also discard infrequent words from our dataset: for each word in a post's main entry, we kept it only if it appears at least one more time in some main entry. We apply the same word pruning to the comment section as well. The corpus size and the vocabulary size of the five datasets are listed in Table 1. In addition, each user's handle is replaced with a unique integer. The dataset is available for download at `http://www.ark.cs.cmu.edu/blog-data`.

3.2 Qualitative Properties of Blogs

We believe that readers' reactions to blog posts are an integral part of blogging activity. Often comments are much more substantial and informative than the post. While circumspective articles limit themselves to allusions or oblique references, readers' comments may point to heart of the matter more

`washingtonmonthly.com` and MY for Think Progress at`http://yglesias.thinkprogress.org`.

boldly. Opinions are expressed more blatantly in comments. Comments may help a human (or automated) reader to understand the post more clearly when the main text is too terse, stylized, or technical.

Although the main entry and its comments are certainly related and at least partially address similar topics, they are markedly different in several ways. First of all, their vocabulary is noticeably different. Comments are more casual, conversational, and full of jargon. They are less carefully edited and therefore contain more misspellings and typographical errors. There is more diversity among comments than within the single-author post, both in style of writing and in what commenters like to talk about. Depending on the subjects covered in a blog post, different types of people are inspired to respond. We believe that analyzing a piece of text based on the reaction it causes among those who read it is a fascinating problem for NLP.

Blog *sites* are also quite distinctive from each other. Their language, discussion topics, and collective political orientations vary greatly. Their volumes also vary; multi-author sites (such as DK, RS) may consistently produce over twenty posts per day, while single-author sites (such as MY, CB) may have a day with only one post. Single author sites also tend to have a much smaller vocabulary and range of interests. The sites are also culturally different in commenting styles; some sites are full of short interjections, while others have longer, more analytical comments. On some sites, users appear to be
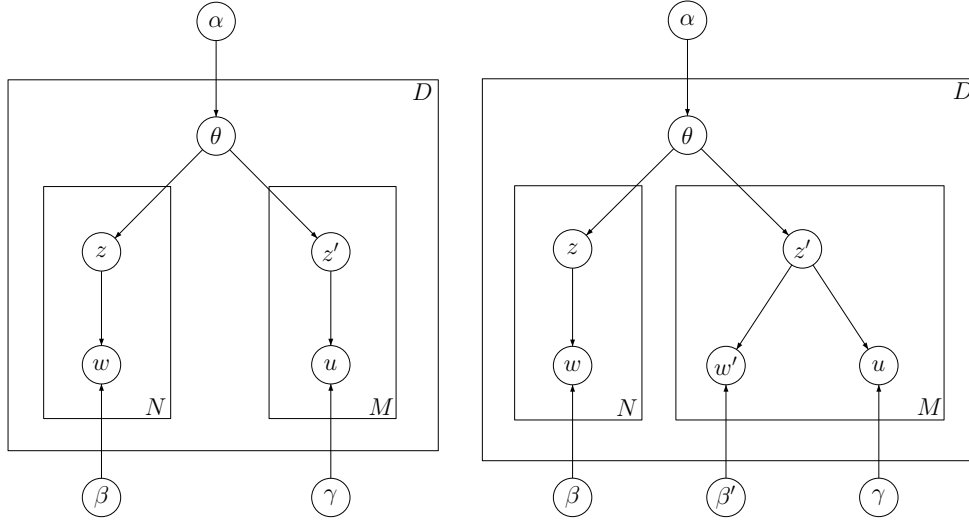
Figure 1: Left: LinkLDA (Erosheva et al., 2004), with variables reassigned. Right: CommentLDA. In training, w, u, and (in CommentLDA) w' are observed. D is the number of blog posts, and N and M are the word counts in the post and the all of its comments, respectively. Here we "count by verbosity."

close-knit, while others have high turnover.

In the next section, we describe how we apply topic models to political blogs, and how these probabilistic models can put to use to make predictions.

4 Generative Models

The first model we consider is **LinkLDA**, which is analogous to the model of Erosheva et al. (2004), though the variables are given different meanings here.[7] The graphical model is depicted in Fig. 1 (left). As in LDA and its many variants, this model postulates a set of latent "topic" variables, where each topic k corresponds to a multinomial distribution β_k over the vocabulary. In addition to generating the words in the post from its topic mixture, this model also generates a bag of users who respond to the post, according to a distribution γ over users given topics. In this model, the topic distribution θ is all that determines the text content of the post and which users will respond to the post.

LinkLDA models which users are likely to respond to a post, but it does not model what they will write. Our new model, **CommentLDA**, generates the contents of the comments (see Fig. 1, right). In order to capture the differences in language style between posts and comments, however, we use a different conditional distribution over comment words given topics, β'. The post text, comment text, and commenter distributions are all interdependent through the (latent) topic distribution θ, and a topic k is defined by:

[7]Instead of blog commenters, they modeled citations.

- A multinomial distribution β_k over post words;
- A multinomial distribution β'_k over comment words; and
- A multinomial distribution γ_k over blog commenters who might react to posts on the topic.

Formally, LinkLDA and CommentLDA generate blog data as follows: For each blog post (1 to D):

1. Choose a distribution θ over topics according to Dirichlet distribution α.

2. For i from 1 to N_i (the length of the post):
   (a) Choose a topic z_i according to θ.
   (b) Choose a word w_i according to the topic's post word distribution β_{z_i}.

3. For j from 1 to M_i (the length of the comments on the post, in words):
   (a) Choose a topic z'_j.
   (b) Choose an author u_j from the topic's commenter distribution $\gamma_{z'_j}$.
   (c) *(CommentLDA only)* Choose a word w'_j according to the topic's comment word distribution $\beta'_{z'_j}$.

4.1 Variations on Counting Users

As described, CommentLDA associates each comment word token with an independent author. In both LinkLDA and CommentLDA, this "**counting by verbosity**" will force γ to give higher probability to users who write longer comments with more

words. We consider two alternative ways to count comments, applicable to both LinkLDA and CommentLDA. These both involve a change to step 3 in the generative process.

Counting by response (replaces step 3): For j from 1 to U_i (the number of users who respond to the post): (a) and (b) as before. (c) *(CommentLDA only)* For ℓ from 1 to $\ell_{i,j}$ (the number of words in u_j's comments), choose w'_ℓ according to the topic's comment word distribution $\beta'_{z'_j}$. This model collapses all comments by a user into a single bag of words on a single topic.[8]

Counting by comments (replaces step 3): For j from 1 to C_i (the number of comments on the post): (a) and (b) as before. (c) *(CommentLDA only)* For ℓ from 1 to $\ell_{i,j}$ (the number of words in comment j), choose w'_ℓ according to the topic's comment word distribution $\beta'_{z'_j}$. Intuitively, each comment has a topic, a user, and a bag of words.

The three variations—counting users by verbosity, response, or comments—correspond to different ways of thinking about topics in political blog discourse. Counting by verbosity will let garrulous users define the topics. Counting by response is more democratic, letting every user who responds to a blog post get an equal vote in determining what the post is about, no matter how much that user says. Counting by comments gives more say to users who engage in the conversation *repeatedly*.

4.2 Implementation

We train our model using empirical Bayesian estimation. Specifically, we fix $\alpha = 0.1$, and we learn the values of word distributions β and β' and user distribution γ by maximizing the likelihood of the training data:

$$p(\boldsymbol{w}, \boldsymbol{w'}, \boldsymbol{u} \mid \alpha, \beta, \beta', \gamma) \qquad (1)$$

(Obviously, β' is not present in the LinkLDA models.) This requires an inference step that marginalizes out the latent variables, θ, z, and z', for which we use Gibbs sampling as implemented by the Hierarchical Bayes Compiler (Daumé, 2007). The Gibbs

---

[8]The counting-by-response models are deficient, since they assume each user will only be chosen once per blog post, though they permit the same user to be chosen repeatedly.

sampling inference algorithm for LDA was first introduced by Griffiths and Steyvers (2004) and has since been used widely.

5 Empirical Evaluation

We adopt a typical NLP "train-and-test" strategy that learns the model parameters on a training dataset consisting of a collection of blog posts and their commenters and comments, then considers an unseen test dataset from a later time period. Many kinds of predictions might be made about the test set and then evaluated against the true comment response. For example, the likelihood of a user to comment on the post, given knowledge of θ can be estimated as:[9]

$$
\begin{aligned}
p(u \mid w_1^N, \gamma, \theta) &= \sum_{z=1}^{K} p(u \mid z, \gamma) p(z \mid w_1^N, \theta) \\
&= \sum_{z=1}^{K} \gamma_{z,u} \cdot \theta_z \qquad (2)
\end{aligned}
$$

The latter is in a sense a "guessing game," a prediction on who is going to comment on a new blog post. A similar task was used by Nallapati and Cohen (2008) for assessing the performance of Link-PLSA-LDA: they predicted the presence or absence of citation links between documents. We report the performance on this prediction task using our six blog topic models (LinkLDA and CommentLDA, with three counting variations each).

Our aim is to explore and compare the effectiveness of the different models in discovering topics that are useful for a practical task. We also give a qualitative analysis of topics learned.

5.1 Comment Prediction

For each political blog, we trained the three variations each of LinkLDA and CommentLDA. Model parameters β, γ, and (in CommentLDA) β' were learned by maximizing likelihood, with Gibbs sampling for inference, as described in §4.2. The number of topics K was fixed at 15.

A simple baseline method makes a post-independent prediction that ranks users by their comment frequency. Since blogs often have a "core constituency" of users who post frequently, this is a

---

[9]Another approach would attempt to integrate out θ.

	$n=5$	$n=10$	$n=20$	$n=30$	oracle
MY					
Freq.	23.93	18.68	14.20	11.65	13.18
NB	25.13	19.28	14.20	11.63	13.54
Link-v	20.10	14.04	11.17	9.23	11.32
Link-r	26.77	18.63	14.64	12.47	14.03
Link-c	25.13	18.85	14.61	11.91	13.84
Com-v	22.84	17.15	12.75	10.69	12.77
Com-r	**27.54**	**20.54**	14.61	12.45	**14.35**
Com-c	22.40	18.50	**14.83**	**12.56**	14.20
Max	*94.75*	*89.89*	*73.63*	*58.76*	*92.60*
RWN					
Freq.	32.56	30.17	22.61	19.7	**27.19**
NB	25.63	**34.86**	**27.61**	**22.03**	18.28
Link-v	28.14	21.06	17.34	14.51	19.81
Link-r	32.92	29.29	22.61	18.96	26.32
Link-c	32.56	27.43	21.15	17.43	25.09
Com-v	29.02	24.07	19.07	16.04	22.71
Com-r	**36.10**	29.64	23.8	19.26	25.97
Com-c	32.03	27.43	19.82	16.25	23.88
Max	*90.97*	*76.46*	*52.56*	*37.05*	*96.16*
CB					
Freq.	33.38	28.84	24.17	20.99	21.63
NB	36.36	31.15	25.08	**21.40**	23.22
Link-v	32.06	26.11	19.79	17.43	18.31
Link-r	**37.02**	31.65	24.62	20.85	22.34
Link-c	36.03	**32.06**	**25.28**	21.10	**23.44**
Com-v	32.39	26.36	20.95	18.26	19.85
Com-r	35.53	29.33	24.33	20.22	22.02
Com-c	33.71	29.25	23.80	19.86	21.68
Max	*99.66*	*98.34*	*88.88*	*72.53*	*95.58*
RS					
Freq.	**25.45**	16.75	11.42	9.62	17.15
NB	22.07	16.01	11.60	9.76	16.50
Link-v	14.63	11.9	9.13	7.76	11.38
Link-r	25.19	**16.92**	**12.14**	**9.82**	**17.98**
Link-c	24.50	16.45	11.49	9.32	16.76
Com-v	14.97	10.51	8.46	7.37	11.3 0
Com-r	15.93	11.42	8.37	6.89	10.97
Com-c	17.57	12.46	8.85	7.34	12.14
Max	*80.77*	*62.98*	*40.95*	*29.03*	*91.86*
DK					
Freq.	24.66	19.08	15.33	13.34	9.64
NB	**35.00**	**27.33**	**22.25**	**19.45**	**13.97**
Link-v	20.58	19.79	15.83	13.88	10.35
Link-r	33.83	27.29	21.39	19.09	13.44
Link-c	28.66	22.16	18.33	16.79	12.60
Com-v	22.16	18.00	16.54	14.45	10.92
Com-r	33.08	25.66	20.66	18.29	12.74
Com-c	26.08	20.91	17.47	15.59	11.82
Max	*100.00*	*100.00*	*100.00*	*99.09*	*98.62*

Table 2: Comment prediction results on 5 blogs. See text.

strong baseline. We also compared to a Naïve Bayes classifier (with word counts in the post's main entry as features). To perform the prediction task with our models, we took the following steps. First, we removed the comment section (both the words and the authorship information) from the test data set. Then, we ran a Gibbs sampler with the partial data, fixing the model parameters to their learned values and the blog post words to their observed values. This gives a posterior topic mixture for each post (θ in the above equations).[10] We then computed each user's comment prediction score for each post as in Eq. 2. Users are ordered by their posterior probabilities. Note that these posteriors have different meanings for different variations:

- When counting by verbosity, the value is the probability that the next (or any) comment word will be generated by the user, given the blog post.
- When counting by response, the value is the probability that the user will respond *at all*, given the blog post. (Intuitively, this approach best matches the task at hand.)
- When counting by comments, the value is the probability that the next (or any) comment will be generated by the user, given the blog post.

We compare our commenter ranking-by-likelihood with the actual commenters in the test set. We report in Tab. 2 the precision (macro-averaged across posts) of our predictions at various cut-offs (n). The oracle column is the precision where it is equal to the recall, equivalent to the situation when the true number of commenters is known. (The performance of random guessing is well below 1% for all sites at cut-off points shown.) "Freq." and "NB" refer to our baseline methods. "Link" refers to LinkLDA and "Com" to CommentLDA. The suffixes denote the counting methods: verbosity ("-v"), response ("-r"), and comments ("-c"). Recall that we considered only the comments by the users seen at least once in the training set, so perfect precision, as well as recall, is impossible when new users comment on a post; the *Max* row shows the maximum performance possible given the set of commenters recognizable from the training data.

---

[10]For a few cases we checked the stability of the sampler and found results varied by less than 1% precision across ten runs.

Our results suggest that, if asked to guess 5 people who would comment on a new post given some site history, we will get 25–37% of them right, depending on the site, given the content of a new post.

We achieved some improvement over both the baseline and Naïve Bayes for some cut-offs on three of the five sites, though the gains were very small for and RS and CB. LinkLDA usually works slightly better than CommentLDA, except for MY, where CommentLDA is stronger, and RS, where CommentLDA is extremely poor. Differences in commenting style are likely to blame: MY has relatively long comments in comparison to RS, as well as DK. MY is the only site where CommentLDA variations consistently outperformed LinkLDA variations, as well as Naïve Bayes classifiers. This suggests that sites with more terse comments may be too sparse to support a rich model like CommentLDA.

In general, counting by response works best, though counting by comments is a close rival in some cases. We observe that counting by response tends to help LinkLDA, which is ignorant of the word contents of the comment, more than it helps CommentLDA. Varying the counting method can bring as much as 10% performance gain.

Each of the models we have tested makes different assumptions about the behavior of commenters. Our results suggest that commenters on different sites behave differently, so that the same modeling assumptions cannot be made universally. In future work, we hope to permit blog-specific properties to be automatically discovered during learning, so that, for example, the comment words can be exploited when they are helpful but assumed independent when they are not. Of course, improved performance might also be obtained with more topics, richer priors over topic distributions, or models that take into account other cues, such as the time of the post, pages it links to, etc. It is also possible that better performance will come from more sophisticated supervised models that do not use topics.

5.2 Qualitative Evaluation

Aside from prediction tasks such as above, the model parameters by themselves can be informative. β defines which words are likely to occur in the post body for a given topic. β' tells which words are likely to appear in the collective response to a partic-

ular topic. Similarity or divergence of the two distributions can tell us about differences in language used by bloggers and their readers. γ expresses users' topic preferences. A pair or group of participants may be seen as "like-minded" if they have similar topic preferences (perhaps useful in collaborative filtering).

Following previous work on LDA and its extensions, we show words most strongly associated with a few topics, arguing that some coherent clusters have been discovered. Table 3 shows topics discovered in MY using CommentLDA (counting by comments). This is the blog site where our models most consistently outperformed the Naïve Bayes classifiers and LinkLDA, therefore we believe the model was a good fit for this dataset.

Since the site is concentrated on American politics, many of the topics look alike. Table 3 shows the most probable words in the posts, comments, and both together for five hand-picked topics that were relatively transparent. The probabilistic scores of those words are computed with the scoring method suggested by Blei and Lafferty (in press).

The model clustered words into topics pertaining to religion and domestic policy (first and last topics in Table 3) quite reasonably. Some of the religion-related words make sense in light of current affairs.[11] Some words in the comment section are slightly off-topic from the issue of religion, such as *dawkins*[12] or *wright*,[13] but are relevant in the context of real-world events. Notice those words rank highly only in the comment section, showing differences between discussion in the post and the comments. This is also noticeable, for example, in the "primary" topic (second in Table 3), where the Republican primary receives more discussion in the main post, and in the "Iraq war" and "energy" topics, where bloggers discuss strategy and commenters

---

[11]Mitt Romney was a candidate for the Republican nomination in 2008 presidential election. He is a member of The Church of Jesus Christ of Latter-Day Saints. Another candidate, Mike Huckabee, is an ordained Southern Baptist minister. Moktada al-Sadr is an Iraqi theologian and political activist, and John Hagee is an influential televangelist.

[12]Richard Dawkins is a well known evolutionary biologist who is a vocal critic of intelligent design.

[13]We believe this is a reference to Rev. Jeremiah Wright of Trinity United Church of Christ, whose inflammatory rhetoric was negatively associated with then-candidate Barack Obama.

483

religion; in both:	people, just, american, church, believe, god, black, jesus, mormon, faith, jews, right, say, mormons, religious, point
in posts:	romney, huckabee, muslim, political, hagee, cabinet, mitt, consider, true, anti, problem, course, views, life, real, speech, moral, answer, jobs, difference, muslims, hardly, going, christianity
in comments:	religion, think, know, really, christian, obama, white, wright, way, said, good, world, science, time, dawkins, human, man, things, fact, years, mean, atheists, blacks, christians
primary; in both:	obama, clinton, mccain, race, win, iowa, delegates, going, people, state, nomination, primary, hillary, election, polls, party, states, voters, campaign, michigan, just
in posts:	huckabee, wins, romney, got, percent, lead, barack, point, majority, ohio, big, victory, strong, pretty, winning, support, primaries, south, rules
in comments:	vote, think, superdelegates, democratic, candidate, pledged, delegate, independents, votes, white, democrats, really, way, caucuses, edwards, florida, supporters, wisconsin, count
Iraq war; in both:	american, iran, just, iraq, people, support, point, country, nuclear, world, power, military, really, government, war, army, right, iraqi, think
in posts:	kind, united, forces, international, presence, political, states, foreign, countries, role, need, making, course, problem, shiite, john, understand, level, idea, security, main
in comments:	israel, sadr, bush, state, way, oil, years, time, going, good, weapons, saddam, know, maliki, want, say, policy, fact, said, shia, troops
energy; in both:	people, just, tax, carbon, think, high, transit, need, live, going, want, problem, way, market, money, income, cost, density
in posts:	idea, public, pretty, course, economic, plan, making, climate, spending, economy, reduce, change, increase, policy, things, stimulus, cuts, low, fi nancial, housing, bad, real
in comments:	taxes, fuel, years, time, rail, oil, cars, car, energy, good, really, lot, point, better, prices, pay, city, know, government, price, work, technology
domestic policy; in both:	people, public, health, care, insurance, college, schools, education, higher, children, think, poor, really, just, kids, want, school, going, better
in posts:	different, things, point, fact, social, work, large, article, getting, inequality, matt, simply, percent, tend, hard, increase, huge, costs, course, policy, happen
in comments:	students, universal, high, good, way, income, money, government, class, problem, pay, americans, private, plan, american, country, immigrants, time, know, taxes, cost

Table 3: The most probable words for some CommentLDA topics (MY).

focus on the tangible (*oil*, *taxes*, *prices*, *weapons*).

While our topic-modeling approach achieves mixed results on the prediction task, we believe it holds promise as a way to understand and summarize the data. Without CommentLDA, we would not be able to easily see the differences noted above in blogger and commenter language. In future work, we plan to explore models with weaker independence assumptions among users, among blog posts over time, and even across blogs. This line of research will permit a more nuanced understanding of language in the blogosphere and in political discourse more generally.

6 Conclusion

In this paper we applied several probabilistic topic models to discourse within political blogs. We in-troduced a novel comment prediction task to assess these models in an objective evaluation with possible practical applications. The results show that predicting political discourse behavior is challenging, in part because of considerable variation in user behavior across different blog sites. Our results show that using topic modeling, we can begin to make reasonable predictions as well as qualitative discoveries about language in blogs.

Acknowledgments

References

L. Adamic and N. Glance. 2005. The political blogosphere and the 2004 U.S. election: Divided they blog. In *Proceedings of the 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*.

D. Blei and M. Jordan. 2003. Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*.

D. Blei and J. Lafferty. In press. Topic models. In A. Srivastava and M. Sahami, editors, *Text Mining: Theory and Applications*. Taylor and Franci.

D. Blei and J. McAuliffe. 2008. Supervised topic models. In *Advances in Neural Information Processing Systems 20*.

D. Blei, A. Ng, and M. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

S. R. K. Branavan, H. Chen, J. Eisenstein, and R. Barzilay. 2008. Learning document-level semantic properties from free-text annotations. In *Proceedings of ACL-08: HLT*.

D. Cohn and T. Hofmann. 2001. The missing link—a probabilistic model of document content and hypertext connectivity. In *Neural Information Processing Systems 13*.

H. Daumé. 2007. HBC: Hierarchical Bayes compiler. http://www.cs.utah.edu/~hal/HBC.

M. Dredze, H. M. Wallach, D. Puller, and F. Pereira. 2008. Generating summary keywords for emails using topics. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*.

E. Erosheva, S. Fienberg, and J. Lafferty. 2004. Mixed membership models of scientific publications. *Proceedings of the National Academy of Sciences*, pages 5220–5227, April.

T. L. Griffiths and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101 Suppl. 1:5228–5235, April.

W.-H. Lin, E. Xing, and A. Hauptmann. 2008. A joint topic and perspective model for ideological discourse. In *Proceedings of 2008 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*.

R. Malouf and T. Mullen. 2007. Graph-based user classification for informal online political discourse. In *Proceedings of the 1st Workshop on Information Credibility on the Web*.

A. McCallum. 1999. Multi-label text classification with a mixture model trained by EM. In *AAAI Workshop on Text Learning*.

T. Mullen and R. Malouf. 2006. A preliminary investigation into sentiment analysis of informal political discourse. In *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*.

R. Nallapati and W. Cohen. 2008. Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs. In *Proceedings of the 2nd International Conference on Weblogs and Social Media*.

M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*.

M. Steyvers and T. Griffiths. 2007. Probabilistic topic models. In T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, editors, *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates.

M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. L. Griffiths. 2004. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*.

I. Titov and R. McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL-08: HLT*.

H. Zhang, B. Qiu, C. L. Giles, H. C. Foley, and J. Yen. 2007. An LDA-based community structure discovery approach for large-scale social networks. In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*.