Predicting Response to Political Blog Posts with Topic Models

Tae Yano
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
taey@cs.cmu.edu

William W. Cohen
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213, USA
wcohen@cs.cmu.edu

Noah A. Smith
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
nasmith@cs.cmu.edu

Abstract

In this paper we model discussions in online political weblogs (blogs). To do this, we extend Latent Dirichlet Allocation, introduced by 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 comment activity on a given post. We also provide a qualitative discussion about what the models discover.

1 Introduction

In recent years web logging (blogging) and its social impact have 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 actions in the blogosphere have 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 arguably 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 (LDA) introduced by Blei et al. (2003). Unsupervised topic models (such as LDA) can be applied collections of unannotated documents, requiring very little corpus engineering. They are also flexible, and can be easily adapted to new problems by altering the graphical model, then applying standard probabilistic inference algorithms for learning and/or prediction (Blei et al., 2003). Different models can be compared to explore the ramifications of different hypotheses about the data. For example, here we will explore whether the contents of posts a user has commented on in the past and the words she has used in her past comments are helpful for predicting 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, most prominently 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 blogosphere. 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 greatest 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 the classical, supervised text classification experiments are mixed (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. This further hinders supervised modeling.

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 . 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) .

Several studies in topic modeling are especially relevant to our work. 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. 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 Griffiths and Steyvers (2004), defined a generative process not only for each word in the text, but also its citation to other documents in the collection, thereby capture the notion of *relation* between the document into one generative process. Nallapati and Cohen (2008) introduced Link-PLSA-LDA model, in which the text contents of the citing document and the "influences" on the document, represented as 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* a single blog conversation, 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 a large collection of 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 foreign 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), Daily Kos (DK), Matthew Yglesias (MY), Red State (RS), and Right Wing News (RWN).

Because our focus in this paper is on blog posts and their comments, we discard posts on which no one commented within six days. All posts and comments are represented as text only (images, hyperlinks, and other non-text contents are ignored). Words occurring two or fewer times and stop words were removed. Posts with fewer than 5 words are discarded. The corpus size and the vocabulary size of the five datasets are listed in Table. 1. Similar preprocessing was done to the comment section of the posts. In addition, each user's handle is replaced with a unique integer.

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 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.

	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	2116	2146
# words (total)	110,788	194,948	183,635	334,051	221,820
(on average per post)	(68.94)	(185.31)	(170.03)	(157.87)	(103.36)
# comments	56,507	34,734	34,244	60,972	425,494
(on average per post)	(35)	(33)	(31)	(28)	(198)
(unique commenters, on average)	(24)	(13)	(24)	(14)	(93)
# words in comments (total)	2,287,843	1,073,726	1,411,363	1,713,505	8,359,456
(on average per post)	(1423.67)	(1020.65)	(1306.817)	(809.78)	(3895.36)
(on average per comment)	(41)	(31)	(41)	(28)	(20)
Post vocabulary size	6,659	9,707	7,579	12,528	10,179
Comment vocabulary size	33,350	22,024	24,702	25,733	58,591
Size of user pool	7,341	963	5,059	2,816	16,849
# test posts	183	143	121	159	240

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

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 enticed 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, CB, RS, and RWN) may produce over twenty posts per day, while single-author sites (such as MY) may average less than one post per day. 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. In some sites, users appear to be 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 answer interesting questions. In doing so, we attempt to capture some of the above mentioned unique characteristics of this social activity.

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.[1] The graphical model is depicted in Fig. 1. 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 not only the text content of the post, but also 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. 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, and a topic k is defined by:

- A multinomial distribution β_k over post words;
- A multinomial distribution β'_k over comment words; and

[1] Instead of blog commenters, they modeled document citations.

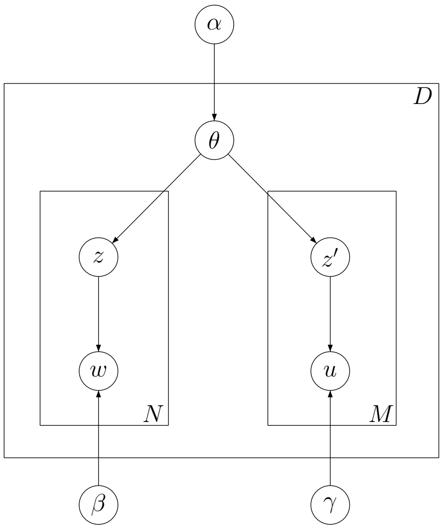

Figure 1: LinkLDA (Erosheva et al., 2004), with variables reassigned for our purposes. In training, w and u are observed. D is the number of blog posts, N is the number of words in the post, and M counts users. See text for discussion of three different ways to count users.

- A multinomial distribution γ_k over blog commenters who might react to posts on the topic.

The graphical model is depicted in Fig. 2.[2]

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):
   (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 $\beta'_{z'_j}$.

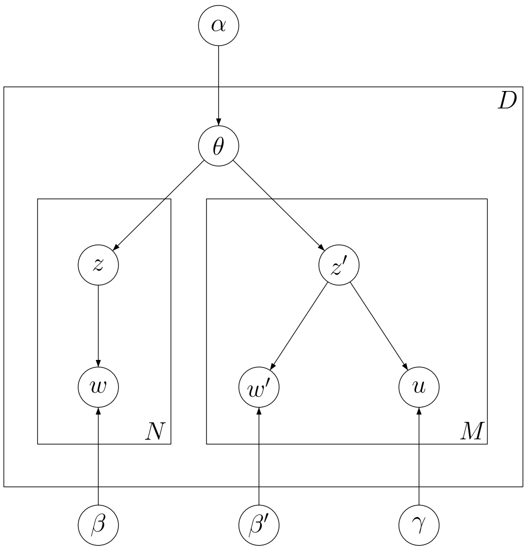

Figure 2: CommentLDA. In training, w, w', and u are observed. D is the number of documents, N is the number of words in the post body, and M is the total number of words in all comments. Here we "count by verbosity." See text for variations.

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.[3]

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 $\beta'_{z'_j}$.

[2] Another model, not explored here, might model the entirety of all comments without modeling the users who generated them. Since our evaluation task (§5) is to predict which *users* will comment on a post, this model did not make sense in our setting.

[3] We note that 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.

This model is perhaps the most intuitive; 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:

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

(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).[4] We also derived and tested mean-field variational inference for some of the models; this achieved similar results (faster) but the implementation was less automated and more error-prone.

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 can be estimated as:

$$\begin{aligned} p(u \mid w_1^N, \alpha, \beta, \beta', \gamma) &= \sum_{z=1}^{K} p(u \mid z) P(z \mid w_1^N) \\ &= \sum_{z=1}^{K} \gamma_{z,u} \cdot \theta_z \end{aligned}$$

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.

As a baseline method we make a static prediction that ranks users by overall comment frequency in the training data. This is a strong baseline, since blogs tend to have a "core constituency" of users who post frequently, regardless of the content of the post.

To perform the prediction task, 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 run 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). We then compute each user's comment prediction score for each post as in Eq. 2. Users are then 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 resopnd *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

[4] `http://www.cs.utah.edu/~hal/HBC`

MY: precision (%) at cutoff n

	n=5	n=10	n=20	n=30	oracle
Base.	23.93	18.68	14.20	11.65	11.42
Link-v	20.10	14.04	11.17	9.23	9.21
Link-r	26.77	18.63	14.64	12.47	11.67
Link-c	25.13	18.85	14.61	11.91	11.21
Com-v	22.84	17.15	12.75	10.69	10.20
Com-r	**27.54**	**20.54**	14.61	12.45	**11.88**
Com-c	22.40	18.50	**14.83**	**12.56**	11.72

RWN: precision (%) at cutoff n

	n=5	n=10	n=20	n=30	oracle
Base.	25.73	23.98	17.93	**15.61**	**16.16**
Link-v	19.30	17.48	13.88	11.95	12.58
Link-r	**27.69**	22.65	**18.11**	15.36	15.45
Link-c	25.17	21.74	16.74	14.26	14.85
Com-v	26.57	20.62	14.82	12.54	14.92
Com-r	25.87	**24.19**	18.04	15.01	15.66
Com-c	25.59	21.46	15.76	13.26	13.63

CB: precision (%) at cutoff n

	n=5	n=10	n=20	n=30	oracle
Base.	33.38	28.84	24.17	20.99	18.78
Link-v	32.06	26.11	19.79	17.43	16.39
Link-r	**37.02**	31.65	24.62	20.85	19.44
Link-c	36.03	**32.06**	**25.28**	**21.10**	**19.82**
Com-v	32.39	26.36	20.95	18.26	17.12
Com-r	35.53	29.33	24.33	20.22	18.77
Com-c	33.71	29.25	23.80	19.86	18.80

RS: precision (%) at cutoff n

	n=5	n=10	n=20	n=30	oracle
Base.	25.50	16.72	10.84	**9.24**	**17.43**
Link-v	13.20	11.25	8.89	7.73	9.14
Link-r	**25.53**	**17.04**	10.84	9.03	17.13
Link-c	24.40	15.78	**11.19**	8.95	16.93
Com-v	13.71	10.37	7.92	6.49	9.72
Com-r	15.47	10.50	7.89	6.75	10.15
Com-c	15.97	11.00	7.76	6.49	10.92

DK: precision (%) at cutoff n

	n=5	n=10	n=20	n=30	oracle
Base.	24.66	19.08	15.33	13.34	7.67
Link-v	20.58	19.79	15.83	13.88	7.69
Link-r	**33.83**	**27.29**	**21.39**	**19.09**	**9.60**
Link-c	28.66	22.16	18.33	16.79	9.03
Com-v	22.16	18.00	16.54	14.45	8.06
Com-r	33.08	25.66	20.66	18.29	9.27
Com-c	26.08	20.91	17.47	15.59	8.78

Table 2: Comment prediction results on five blogs. "Link" refers to LinkLDA and "Com" to CommentLDA. The suffixes denote the counting methods: verbosity ("-v"), response ("-r"), and comments ("-c"). "Base." refers to our baseline method.

in Tab. 2 the precision (macro-averaged across posts) of our predictions at various cut-offs. The oracle is the precision where it is equal to the recall, equivalent to the situation when the actual number of commenters is known.

As noted, we considered only the comments by the users seen at least once in the training set, so perfect recall is impossible when new users comment on a post. Perfect precision is very difficult to achieve, except at meaninglessly low thresholds, since many effects are not captured in our models (e.g., the time of the posting relative to a user's waking or reading hours). The performance of random guessing is well below 1% for all sites at cut-off points shown.

We achieved some improvement over the baseline for small cut-offs on all five sites, though the gains were very small for RWN and RS.[5]

Our results suggest that if we are 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 and the volume of the response. The task naturally is more difficult when the user set is large. DK, with 93 participants per post on average, had the lowest performance at the oracle precision. We attribute this result to the inherent difficulty in this site since our baseline performance is also quite low at that cut-off. Interestingly, in this site, our best model brought the largest gain over the baseline at all the cut-off points.

LinkLDA usually works slightly better than CommentLDA, except for MY, where CommentLDA is stronger, and RS, where CommentLDA is extremely poor. Again, differences in commenting style are likely to blame: MY has relatively long comments, and RS has the shortest average total comment length.

In general, counting by response works best, though counting by comments is a close rival in some cases. Varying the counting method can bring as much as 10% performance gain. We observe that counting by response helps LinkLDA, which is ignorant of the word contents of the comment, more than CommentLDA. As a consequence, counting by response helps more at the sites where LinkLDA does better. In those sites, the performance of LinkLda was often worse than CommentLDA under counting by verbosity or comments.

[5] We note that these are the more conservative blogs, hinting at a difference in commenting styles that correlates with ideology. Further exploration is required to test this idea.

Closer inspection into the sites' profiles is revealing. Both MY and CB have larger average comment words per commenter compared to the main contents of the posts. It is more so in MY, which is the only site where CommentLDA variations consistently outperformed LinkLDA variations. This suggests that if the site, on average, contains less verbose comments, ignoring the difference in each comment (by not counting by verbosity, or by choosing LinkLDA) actually helps with predicting who is going to comment on the post. On the other hand, if participants are more expressive, the discrimination based on the comment contents will help the prediction task more.

5.2 Qualitative Evaluation

Aside from the prediction tasks such as above, the model parameters by themselves can be informative. β' tells which words are likely to appear in the collective response to a particular topic. Similarity or divergence of the two distribution given the topic can be useful in analyzing the the reaction to the post. Parameter 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). β defines which words are likely to occur in the post body for a given topic. Following previous work on LDA and 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 DK data using CommentLDA with counting by verbosity and $K = 20$.[6] The model is trained slightly differently from those reported in §5.1. More aggressive pruning was applied, mainly to force the resulting model parameters to be more understandable. Specifically, users who wrote fewer than 100 words were removed, leaving 3,612 users and a comment vocabulary of 37,976 words (over a half million words remain in the comments).

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 four hand-picked topics that were relatively transparent. Topic 2 corresponds to science, global climate change in particular. Topic 12 clearly is on the racial aspect of the Democratic presidential campaign.[7] Without some knowledge of current affairs, the connection between these words is, of course, not obvious. Notice that the actual term "racism" appeared only in the comment text, reflecting its frankness and suggesting that bloggers on this site shy away from using the word directly even when it is central to the discussion.

In topic 7, on the Democratic primary in Iowa, though other candidates' names were frequently mentioned in the post body itself, the popular reaction in the comments seems to concentrate on only Clinton, Obama, and Edwards. The observation is accentuated by a Republican candidate (McCain) who was mentioned more than the bottom candidates. The same tendency is seen in Topic 11, on the CNN debate among the Democratic candidates, of which the comments named Richardson and Dodd less frequently than the post.

Such comparison of words across the document collections is only possible with CommentLDA (not LinkLDA), which jointly learns about posts and comments, though with different distributions. Through our model, the two different realizations of each topic are separated, allowing more nuanced inspection. One difference in the two styles that is easily apparent is the more subjective nature of the comments. More subjective words appeared in the comments, which in general makes it difficult to topically classify the text, insofar as subjective words cut across topics. Our model, however, is capable of associating such writing to purported topics via its association to the posts, which contain better cue words.

Future work might extend the models to learn these properties and make predictions accordingly, or to predict blog response *content* (not just who will comment), or to break temporal and inter-user independence assumptions made by LinkLDA and CommentLDA. We might also consider models that predict discourse in more than one blog at a time.

6 Conclusion

In this paper we applied several probabilistic models (all variations on LDA) to discourse within a political blog. We introduced a novel comment prediction task with

[6] This is not the strongest of our models, but it was the earliest, and these results were generated *before* the other counting methods were developed.

[7] We believe *wright* refers to Rev. Jeremiah Wright of Trinity United Church of Christ, whose inflammatory rhetoric was negatively associated with presidential candidate Barack Obama. The word *ferraro* likely refers to Clinton supporter Geraldine Ferraro's remarks on Barack Obama's credential was widely criticized as racist, and she later become a Fox news contributor.

Topic 2	**“environment and science”**
in posts	climate, news, oil, universe, scientists, water, park, place, today, old, research, young, ago, record, america, ice, environmental, cell
in comments	think, know, need, really, power, say, want, work, things,nuclear, god, point, problem, life, better, solar, believe, far
in both	science, global, just, going, year, time, warming, change, big, world, good, people, energy, way, earth, long, day, little, lot, years, right, thing
Topic 7	**“Iowa caucus”**
in posts	poll, hampshire, supporters, caucus, dodd, public, numbers, results, big, political, polling, second, lieberman, huckabee, richardson, kucinich, today
in comments	think, right, going, really, way, good, say, party, want, state, democrats, kos, election, said, war, mccain, president
in both	obama, edwards, iowa, campaign, candidates, hillary, clinton, people, win, polls, just, point, vote, voters, democratic, thing, nh, know, support, primary, candidate, time, money
Topic 11	**“CNN Democratic candidate debate”**
in posts	post, dodd, democratic, change, policy, richardson, says, america, iraq, biden, republicans, comments, iran, night, important
in comments	really, going, said, say, vote, support, got, years, media, thing, point, saying, lot, great, look
in both	obama, question, hillary, clinton, debate, edwards, bush, campaign, people, want, candidate, just, things, candidates, think, right, need, time, better, war,good, know, president, cnn, way
Topic 12	**“racial issues”**
in posts	barack, ferraro, state, states, john, youtube, moment, real, political, fox, year, got, video, days, sure, despite, saying
in comments	time, way, right, say, going, vote, want, candidate, president, point, america, racist, need, things, man, democratic, american
in both	obama, clinton, people, years, said, campaign, mccain, just, speech, wright, black, war, white, party, race, hillary, good, media, think, know, thing, country, really

Table 3: The most probable words for some topics in posts and comments.

those models to assess their fitness in an objective evaluation with possible practical applications. The results show that using topic modeling, we can begin to make reasonable predictions on a very difficult task. They also show considerable variation in what works best on different blog sites with different properties.

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, 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*, pages 263–271, Columbus, Ohio.

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

Hal Daumé. 2007. HBC: Hierarchical Bayes compiler.

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.

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*.

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 UAI*, Arlington, VA, USA.

M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. L. Griffiths. 2004. Probabilistic author-topic models for information discovery. In *Proceedings of KDD*.

I. Titov and R. McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL-08: HLT*, pages 308–316, Columbus, Ohio. Association for Computational Linguistics.

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 *In Proceedings of the IEEE International Conference on Intelligence and Security Informatics*.