

**Text as Actuator:
Text-Driven Response Modeling and Prediction in Politics**

Tae Yano
taey@cs.cmu.edu

September 5, 2011

Contents

1	Statement of Purpose	1
1.1	Text and Response Prediction	1
1.2	Our Proposed Prediction Tasks	2
1.3	Thesis Statement	3
1.3.1	Contributions	3
1.4	Road Map	4
2	Completed Work: Predicting Responses in Political Blogs	5
2.1	Background: The Political Blogosphere	5
2.2	Task	7
2.3	Political Blog Corpus	7
2.4	Basic Model	8
2.4.1	Approach	8
2.4.2	Summary of Latent Dirichlet Allocation	9
2.4.3	Accounting for user comments	10
2.5	Prediction Models	11
2.5.1	Predicting interested users	11
2.5.2	Results from user prediction	13
2.5.3	Predicting post popularity	15
2.5.4	Results from popularity prediction	16
2.5.5	Summary of results	18
2.6	Related works	19
2.7	Conclusion and future works	21
3	Proposed Works	22
3.1	Surviving Committee: Predicting Bill Success	22
3.1.1	Background	22
3.1.2	Congressional Committee Data	24
3.1.3	Task	26
3.1.4	Proposed works	26
3.1.5	Conclusion and future work	30
3.2	Motivated Voters: Predicting Polling Decision	31
3.2.1	Background	31
3.2.2	DPTE: Dynamic Process Tracing Environment	33
3.2.3	Task	34
3.2.4	Proposed works	34
3.2.5	Conclusion and future work	38
4	Timeline	40

Read Me First

The goal of this document is to communicate the intents; What we are going to do, why they are important, and how we are going to achieve them within the proposed timeframe.

The first part of the proposal covers thesis statement, motivations, and contributions. The thesis statement identifies three prediction tasks. In the second half of the proposal we will discuss those tasks in turn. Each of the three tasks is self-contained and should be able to read independently in any order.

Chapter 1

Statement of Purpose

We will develop a series of prediction tasks on *actuating* text in this work. In our context, actuating text is a text which evokes, or is written to evoke, *responses* from its readership. Many types of online document collections fit this description. Examples include blog posts with readership comments, product reviews with social tagging or ratings, and news stories amplified and spread by quoting or forwarding. Some long-existing corpora, such as the congressional bill collection or floor debate manuscripts, can be seen as variations of this type of text, as voting results or amendments are, in a sense, a collective reaction from the legislative body to the bill or deliberation.¹

The main task of this dissertation work is to deliver novel *prediction* models on responses based on statistical analyses of the associated texts. We utilize *latent variables* to effectively relate documents to the evoked response by learning from data.

The interactions between texts and response could reveal a variety of interesting social meanings. Building models which infer the response from the text is one way of inquiring into the nature of those interactions. In this dissertation, we will consider three distinctive kinds of document collections, each with the novel prediction tasks related to *politics* in the United States.

1.1 Text and Response Prediction

Why should we care about predicting response from text? First, community-oriented documents such as those mentioned above are becoming more and more prevalent, and there are many practical problems concerning these documents. Additionally, many types of user-generated content, often text, are increasingly the subjects of research works in sentiment analysis or knowledge discovery. [70, 96, 8, 89] Moreover, since many of those texts are byproducts of fast-growing modes of public interactions, they are often studied by the social science researchers interested in collective human behavior and its dynamics.[25, 49, 101]

Notice that a broad range of pragmatic questions in this domain can be casted as a form of “response prediction”. Consider the case of a lazy blog reader who dislikes wasting time with boring news, and suppose he wishes to read only the most popular blog posts among the hundreds. There are potentially many ways to define the “popularity” of a writing, but one straightforward approach is perhaps to use the readership responses as a proxy for a popularity measure. The reader therefore wishes to find an article which gathered many responses from the readership, or, better yet, *will* gather many responses in the future. If one contemplates a document recommendation system for users who value *recency*, predictive mechanisms which give a reasonable prediction of the response volume would certainly be

¹To be sure, as we defined above, actuating texts does NOT have to be user-generated contents or of social media, although they are perhaps the most visible examples today. The increased visibility of those media is certainly one factor to motivate response predictions such as ours here.

desirable.

Consider further the situation when the reader wants advice on what *would be* interesting to *him*. This is a question often raised in personalized recommendation systems. At the core of any such system is some strategy for predicting the user’s future response (whether or not they will find it interesting) to the text. Note that this setting arises in many types of document collections where there is a large volume of texts (e.g., news feeds, conference papers, peer reviews on movie or products, tweets) and the system is served to the personal information needs of the user. Similar questions are also posed as strategic text placement problems. For online media or advertisement, selecting “the most important texts” to serve to users’ limited attention or browsing space is a critical business decision.

A few other natural language processing tasks are essentially a type of response prediction. Joshi et al. [47] presented text-driven movie revenue prediction tasks. Their model seeks to predict the moviegoers’ box-office spending from the reviews written by movie critiques. The underlying assumption is that moviegoers’ actions are somehow influenced by the reviews. Gerrish and Blie [34] examined the prediction of congressional action from the bill texts. The same authors also addressed the citation patterns in scientific paper collections. [33]. Citations are in a sense a type of readers’ response, indicative of interests or agreement toward the target publication. Yogatama et al. [103], also [24], used a much different approach to address this same question. Some types of document-level sentiment prediction tasks seek to predict a binary response (“thumbs up”) or a numerical response (such as star rating) from the readership based on the movie review or product description. The question can be cast as a prediction of user reaction caused by the document contents. [74]

1.2 Our Proposed Prediction Tasks

In this work, we present three case studies of text-driven prediction in the domain of American politics and governments:

- Predicting popularity of political blog posts, within a community, given a blog post: models for comment generation in political discussion forums.
- Predicting committee decisions given a congressional bill: models for bill survival through the congressional committee system.
- Predicting an individual’s voting decision change given a trace of articles viewed by the person: models for personal voting behavior during campaign media coverage.

For convenience, we will always refer to the reactions from the readership (in a variety of forms) as the “response”, or “response variable”, in this dissertation. We will call the textual data which is associated with the response the “document” or “actuating document” when it is not clear from the context. Actuating documents always precede the response. The working theme of these tasks is to design stochastic models of (actuating) documents coupled with their responses.

We will train these predictive models in a supervised learning setting. The models will learn the patterns between the document and response from the paired examples in the training data. At the test time, the model is to predict unseen responses from a new document. Here are some more general settings we will assume throughout the rest of this work.

We take it for granted that the two components (documents and their responses) are given, well defined, and presumably interdependent. We assume that the detail of the linkage between the two components is not explicit. Even when there are seemingly apparent links, more useful and better generalizable structures may be latent. For example, given a text and a group of people who responded to the text, we do not necessarily know what elements of the text captured the attention of each person. Furthermore, it is possible that some of the respondents reacted to different

elements of the text from others, and perhaps for different reasons. We presume that annotating all these detailed analyses is expensive, or else are not well defined, and wish to avoid such cost. To overcome such constraints, we advocate a data-driven approach. Instead of adding human annotations on the linkage detail, we will encode them as a set of latent variables in the models. The details of those linkage structures, naturally, will differ for each question.

1.3 Thesis Statement

In this dissertation, we develop a set of novel statistical models for predicting response actuated by text. We examine three types of response related to American politics: reader responses to political blog posts; congressional committee reactions to a proposed piece of legislation; and voter preferences in the presence of campaign media coverage. For each task, our goals are to (1) build models which yield high prediction accuracy and (2) provide a human-understandable data-driven explanation of the underlying response process. We employ latent variable models to address these goals.

Our chosen tasks deal with interesting and relevant subject matters in contemporary politics. Progress in this area is of high concern to social scientists and political scientists, and also offers novel contributions to statistical text analysis research. We anticipate that models like the ones we introduce will ultimately be useful in applications like recommendation and filtering systems, as well as in social science research that makes use of text as data; development of such applications is outside the scope of this thesis.

1.3.1 Contributions

In the beginning of this chapter, we motivated response predictions from the point of practical applicability. In this section, we will note our other contributions in other contexts. The discussion here is applicable to all three prediction tasks. We will note the task-specific contributions later in each chapter.

Statistical analysis of text for extrinsic prediction tasks is a subject that has been explored before, but it is only recently that the field has started to receive a steady stream of attention from the natural language processing research community. Text-based analysis of reader reactions are dealt with in such areas as sentiment analysis, opinion mining, and most recently, text-driven forecasting. Our response prediction models are novel contributions to these growing fields of natural language processing research.

The essence of text-driven forecasting tasks is the exploitation of textual evidence to predict real world events. In a closely related area, an increasing number of quantitative political scientists advocate “text-as-data” [54, 55] approaches to various problems. The key idea in this approach is to treat text as another categorical data in the statistical analysis. Similar algorithms are used in both text-driven forecasting and text-as-data approaches to political science, but their emphases are slightly different. Political scientists are more interested in the explanatory power of statistical models (for example, how meaningful they capture and represent the signals in the text), while text-driven forecasting tends to care more about quantitative predictive performance. As our work holds much relevance to both disciplines, we maintain both of those goals. We hope our work is a meaningful contribution from both perspectives.

In this work, we include experiments based on latent variable modeling in all our tasks. The approach is studied extensively in language processing tasks, such as part of speech tagging, and semantic parsing [37, 79]. Its use in the fields in question is much less explored. However, we believe our response prediction models have much to benefit from this approach. In this work, we demonstrate this point by showing its usefulness in concrete examples.

Latent variable modeling offers a flexible means to express structures among the various evidence without overly

committing the minute details. In the case of our response prediction modeling, it allows us to incorporate the social and political science insights in our predictive model naturally. In addition, the examination of latent variables (along with model parameters) *learned* from the data could yield new insights into the social relations represented in the texts and their reactions, potentially adding useful exploratory power beyond the performance accuracy. Topic models, a type of latent variable model on textual data, have been used in various tasks in relation to corpus analysis in political science largely for such purposes.[67, 38, 83]

1.4 Road Map

We will describe each prediction task in more detail in the rest of the paper. Each of the next three sections corresponds to one of our three focus tasks. The structure of each is essentially parallel.

We first describe the background of our task and the corpora. All our corpora are closely related to some interesting subjects in current politics. We will discuss the significance of these texts, both in real life and in academic research. Those discussions will motivate our particular model design choice. We will also report our exploratory data analysis on the corpus.

We then present the specification of basic models and some extensions. For those previously published works we will present some excerpt from the model evaluation. For those in the earlier stages of development, we will include the development process. We conclude the chapter with a plan for future development.

Chapter 2

Completed Work: Predicting Responses in Political Blogs

In this section we describe our first task, modeling of posts and user responses in political blogs. The goal of our models is to answer whether a given blog post will prompt response from the readerships in the form of commenting. The prediction setting is quite straightforward; A predictor will take a new blog post, then output a prediction on some properties (attributes) of the would-be response. Our task is to design such predictive models, implement them, and experiment on them with real world data.

We think this is quite practical work, since blogging, though it is a relatively new mode of journalism, plays an increasingly important function in informing people of contemporary politics [98]. A mechanism which can forecast how people would react to each post could serve as a core analytic tool for recommendation or filtering system for the readers.

Aside from the practical utility, exploring a blogging community by modeling its response pattern can be a thought provoking exercise. Political blog sites typically form around a groups of people with similar ideology [56, 48], Presuming these communities each have distinctive attitudes toward various issues, and these unique attitudes can be captured in our models, comparing (profiling) them by how those models differ could be an interesting way to describe their political subcultures.

We will start with a discussion on political blogging. We then present our general approach, and gradually extend the idea to the prediction tasks. We have the two types of user response predictions: 1) Predicting which readers would comment on the post, and 2) Predicting the total size of the response. The chapter will conclude with some results from those experiments and the plans for future works.

This chapter describes the works previously published. See [99] and [100] for the full reports.

2.1 Background: The Political Blogosphere

Blogging is studied by computer scientists who research large scale networks (both social and otherwise) or online communities [57, 3, 58, 39]. Among natural language processing researchers, bloggings or other user generated texts are particularly important for sentiment analysis or opinion mining [72, 7, 17, 52, 36].

Blogging is also an important subject in political science [98, 48, 68, 60]. As a means of political journalism, blogging

differs from the traditional main stream media (MSM) in many ways. One of the most apparent differences is bloggers' carefree attitude toward neutrality and objectivity. This is not to say that all blogging is by definition subjective, but rather, that for most of bloggers neutrality is just another option. In main stream journalism objective reporting is a social contract. ¹ Meanwhile, bloggers do not abide to such expectations; They are *supposed* to be subjective, personal, and emotional. This culture of freedom of subjective speech positions blogging as an ideal thought outlet for the concerned citizens.

Blogging however is far more important than a soapbox in our time [98]. It is a means to cultivate a **community** of likeminded people, providing a permanent address where opinions can flourish. The development of community is perhaps in part due to the two-way communication infrastructure that is built into blogging interfaces, the readers' ability to "chime in" his/her thoughts as comments. The infrastructure, combined with aforementioned editorial freedom for partiality, seems to make blogging a media of grassroots or fringe politics. ²

Another trait in political blogging that sharply contrasts with main stream journalism is its unabashed **partisanship** [56]. Unlike the MSM, many of the popular blogs such as Daily Kos³, Think Progress⁴, Hot Air⁵, or Red State⁶, are not only more opinionated, but also unyieldingly partisan.

Related, or perhaps a consequence of this partisan culture is an apparent **balkanization** of blog journalism. In their seminal study of the political blogosphere, [1], and also [56, 48], argued that the political blogosphere is an uncompromisingly divided world. They found that blogging communities prefer to form ideologically homogeneous subgroups, rather than reaching out to the other side of political spectrum. Other studies on the blogosphere observe its echo chamber effects [35], which likely reinforce partisan view points.

As a consequence of this populism, partisanship, and balkanization, the political blogosphere comes to be rather a unique microcosm of contemporary community politics. In this sense, the political blogosphere presents itself as an unprecedented research opportunity; what can we find in this huge quantity of spontaneous, near-realtime trace of political thought and behavior, which likely mirrors various political subcultures in real life?

Motivation

Why is looking into readership response important here? One reason is simply the lack of precedent. Popular blog sites often generates a large volume of user texts in the form of comments. Often the total volume of comments is a magnitude larger than the volume of posts themselves. Also, user comments are potentially useful for sentiment analysis or opinion mining. Mishne and Glance [66] showed the value of comments in characterizing the social repercussions of a post, including popularity and controversy. Nonetheless, user comment contents (textual or otherwise) have so far been rather neglected in the study of user generated texts, including in blogs [100]. ⁷ Part of the motivation in this task is to fill in this gap.

Another way to look at this work is as an investigation into the use of user generated texts in illustrating their communities. Feedbacks from the engaged readers must to some extent define the culture of the site. One idea is that, since blog posts and user comments form a stimulus-response relationship, comments define the community by shaping the interactive patterns between the texts (blog posts) and reader response (comments). We think that building prediction models of response is one way to elucidate such relationships. For example, depending on the ideological orientation

¹Although some may disagree with this point, main stream journalism does at least make attempts to *appear* impartial.

²In many cases, a political blog has more function than a discussion forum. They are often used as a means of activism, such as a call for boycotts or petition or donation or participation for political events such as rallies or demonstrations. See [98] for more studies on the type of political blogging activities.

³<http://dailykos.com/>

⁴<http://thinkprogress.org/>

⁵<http://hotair.com>

⁶<http://www.redstate.com/>

⁷Although we saw more research works on user comments in last couple of years. Our initial blog comment works are cited by some of them [75, 31].

of community, certain issues stimulate much response, while others are ignored by the readers. In a way, a sensible response prediction models *ought to* latch on to such characteristics in community. Conversely, a prediction model which yields higher accuracy would most likely render its discriminative power from knowing such trends.

2.2 Task

As a part of this dissertation work we present statistical models for predicting reader responses to posts in a given political blog community. The models are to output a prediction about some properties (attributes) of the would-be response given a previously unseen blog post. Throughout this section we consider user comments as the sole representative of the reader response. The identity of the blog site is known in advance.

At the prediction time the only contents of the post that are available to the predictor are the words in the post's main entry. No information on the post's context (such as the date, time of the day, bylines, or previous threads) are present. Also, the models are not given any information on any part of the user comments (such as words in comments, user handles, or volume counts) of the current target post. However, the models may use historical comment data for the purpose of training.⁸

We introduce two types of response prediction models. The first one is for commenter handle prediction. This model is to predict a set of registered user handles who comment on the target post. We evaluate this model against the held-out test set, by the accuracy, precision, and recall on how many of the actual commenters are in the predicted set. The second model is for comment volume prediction. In this case, the model is to predict either the count of words or the count of commenters. We evaluate the model by the accuracy, precision, and recall of binary prediction on whether or not the volume of comments is more than the corpus average.

In all our experiments we trained and tested the model with the blog corpus prepared by our team. See table 2.1. We use 10 percent of the data for the held-out evaluation.⁹

2.3 Political Blog Corpus

To support a data driven approach in political blogging, we have collected blog posts and comments from 40 blog sites focusing on American politics during the period from November 2007 to October 2008, contemporaneous with the United States 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 the period and use them to experiment with our prediction models. In this proposal we will discuss only the experiments runs on MY (Matthew Yglesias) and RS (Red State) data set.

See table 2.1 for the detail of this data. The data is available from <http://www.ark.cs.cmu.edu/blog-data/>. The data have been used in several publications to date, such as [6, 4].

⁸Post contents, for this work at least, consist only of texts. We ignore all the non-textual content (such as sounds, movie, graphic, or link to the other pages) at any time.

⁹The resource is available for public use in <http://www.ark.cs.cmu.edu/blog-data/>

	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)
(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 2.1: Details of the blog data used in this paper. “MY” = Matthew Yglesias, “RWN” = Right Wing News, “CB” = Carpet bagger, “RS” = Red State, “DK” = Dairy Kos.

2.4 Basic Model

In this section we will first discuss our approach, then a basic model. We will extend this model for our stated prediction tasks in the subsequent section.

2.4.1 Approach

We wish to design response prediction methods which exploit community-specific response trends. The difficulty, of course, is that such concept as “response trends” is usually ill-defined, and difficult to codify directly. As we laid out in the introductory chapter, we approach this predicament with a data-driven attitude; We encode our uncertainty as a set of latent variables in a probabilistic system, then learn those variables from the data.

In this particular task we approach the response prediction as a case of generative text modeling. This means that we assume a probabilistic generative story over the observable texts. In addition, we also assume that the readership response is a part of this generative process. Once the model is trained, we can simply form the response prediction procedure as a standard probabilistic inference within this model.

One advantage of generative approach is that it is easy to spell out the probabilistic hypothesis, and ties them into actionable computational frameworks.¹⁰ Here are some hypothesis we like to encode into our model: We would like to say that the model jointly generates main text and the responses, and the two components are somehow linked. We also like to express that the contents of the blog post is driven by the set of thematic issues chosen by the authors. and those themes are also shared by the response from users, thereby the same set of underlying concept dictate the response contents. However, the surface expression of those concepts could be much different from the main entry.

There are many approaches with which we can formalize those hypothesis in a generative framework. In this task we use latent topic modeling [93]. The approach is often used in natural language processing research for document clustering or dimensionality reductions. The particular instance of topic model we start with is Latent Dirichlet Allocation (LDA) [13]. LDA, like many other topic models, embodies the idea that the text generation are driven by a set of (unobserved) thematic concepts, and each documents are defined by a subset of those concepts. Due to its

¹⁰The other advantage is that, in this way, we can take advantage of the rich body of inference techniques research in Bayesian statistics.

flexibility, LDA has been extended to many type of statistical text analysis tasks. ¹¹ In our model we extend this idea to include the generation of response values.

2.4.2 Summary of Latent Dirichlet Allocation

Latent Dirichlet Allocation can be described as a model of topic detection over documents, but in our context it is perhaps more instructive to view the model as a Bayesian extension to the class-specific language model or the 0th order Markov model over the text.

Let w_d as a document represented as a bag of unigrams w , and z_d as the document's word class, which has an associated (class conditional) unigram language model $p(w|z)$. Then, one way to express the likelihood function for the word and the class is the following:

$$p(w_d, z_d) = p(z_d; \theta) \cdot \prod_n^{N_d} p(w_{d,n} | z_d)$$

We assume here that the texts are represented as multinomial distribution(s) over the finite vocabulary, although in this equation this notion is not explicit. Assuming multinomial distribution, the parameters for those models can be estimated via maximum likelihood estimate quite easily when all the document-class labels are observed. In English, the above function express the following generative story. Note that this represent a stochastic system equivalent to the one represented in the function above:

1. Choose a class label z_d according to a parameter θ .
2. For n from 1 to N_d (the length of the document):
 - (a) Choose a word $w_{d,n}$ according to the class's word distribution $Multi(w|z)$.

When the class variables are observed, this is the form of joint distribution which Naive Bayes classifiers are derived from. Naive Bayes is used extensively for document classifications in supervised setting.

There are three concepts which LDA adds to this simple model. 1) Each word can be associated with different classes. In LDA, the thematic classes are the "topic". 2) Each topic class is a draw from a document-specific multinomial distribution over the topic. Note that, up to here, LDA's assumption is the same as multinomial mixture models [13]. The last assumption which makes LDA distinct is 3) The document-specific multinomial distribution is also a draw from some distribution. This distribution is a Dirichlet distribution in LDA. In general, this does not have to be a Dirichlet distribution, although it is a convenient choice. This makes inference procedure much simpler since it is a conjugate prior to the multinomial distribution. This approach, treating a model parameters (i.e., the multinomial distribution parameter) as another set of random variables, is often called Bayesian approach. All in all, this come to represent much different generative story from the above. For each document d from 1 to D :

1. Choose a distribution θ_d over topics according to a symmetric Dirichlet distribution parameterized by α).
2. For n from 1 to N_d (the length of the document):
 - (a) Choose a topic $z_{d,n}$ according to the topic distribution θ_d .
 - (b) Choose a word $w_{d,n}$ according to the class's word distribution $\beta_{z_{d,n}}$.

¹¹LDA is a formalism applicable to any type of categorical data. Its use is by no mean limited to textual data, nor to natural language research.

In above notation, we represent the word probability distribution as parameterized by a (topic by vocabulary) matrix β for the readability. Corresponding likelihood function is the following:

$$p(\mathbf{w}_d, \mathbf{z}_d, \theta_d) = p(\theta_d; \alpha) \cdot \prod_n^{N_d} p(w_{d,n}; \beta_{z_{d,n}}) \cdot p(z_{d,n} | \theta_d)$$

In many algorithms for parameter estimation with compound stochastic models such as LDA, inference of the posterior distributions over hidden variables are required. In our experiment (2.5) we employ a stochastic version of Expectation Maximization approach for model training, during which we calibrate the posterior distributions explicitly.^{12 13}

2.4.3 Accounting for user comments

There are a few properties in political blogging we wish to capture, but are not quite represented in LDA. LDA assumes that the generation of text is driven by a set of topics or issues specific to each document. This assumption is fairly reasonable with political blogging since discussions in politics are issue-oriented in nature. If we view a blog document, main entry and comments section all inclusive, as one piece, plain LDA is sufficient. However, this view does not capture the qualitative difference between the languages in post and comment. First of all, their vocabularies are noticeably different. Comments are usually more casual, conversational, and full of jargons.¹⁴ The comment section is also more diverse in styles, for they are written by many more authors than the main entries.

We therefore assume here that the two parts share the same topic set, but the surface expression of those topics are much different, much like two documents each written in different languages. One way to modify the basic LDA to reflect this idea is adding another set of vocabulary to the model: For each document d from 1 to D , we add an additional generative step as:

3. For m from 1 to M_d (the length of the comments on the post, in words):
 - (a) Choose a topic $z'_{d,m}$ according to the distribution θ_d .
 - (b) Choose a word $w'_{d,m}$ according to the comment word distribution $\beta'_{z'_{d,m}}$.

In effect, here we claim that two different vocabularies describe the same set of topics. Note that an additional set of conditional distribution over words, β' , is used for comment side. The arrangement is to capture the differences in language style between posts and comments. Has this make inference (and prediction) algorithm any more difficult? So far not; The only change in the likelihood function is an extra summation term for the comment side:

$$p(\mathbf{w}_d, \mathbf{w}'_d, \mathbf{z}_d, \mathbf{z}'_d, \theta_d) = p(\theta_d; \alpha) \cdot \prod_n^{N_d} p(w_{d,n}; \beta_{z_{d,n}}) \cdot p(z_{d,n} | \theta_d) \\ \cdot \prod_m^{M_d} p(w'_{d,m}; \beta'_{z'_{d,m}}) \cdot p(z'_{d,m} | \theta_d)$$

¹²This is not the only way, In some algorithms, inference over posterior distribution is more implicitly. Once we have estimated model parameters, we can conduct prediction over a new document fairly easily. There are variety of algorithms which could solve the inference problem in stochastic system like LDA. [11, 93, 44]

¹³We do not spell out the full derivation of inference steps for those models and other subsequent models in this proposal. See [99] and [100] for full detail. For general guidelines on how to conduct derivation of compound categorical distributions, the best practical discussion on would be found in [44].

¹⁴They are less carefully edited and therefore contain more misspellings and typographical errors.

2.5 Prediction Models

The blog model in the previous section is essentially a language model, which could reason about the comment content from the main post. In this section we will build upon this basic model, augmenting it with the response variables of our interest. Once those response elements are explicit in the model, we can form prediction on them in straightforward manner using standard probabilistic inference techniques. This type of approach to predictions has been studied in such works as [12, 84, 10].

In this section we will develop two prediction models. For each model, we first discuss the design of the model, then reports the model evaluation on real world data.

2.5.1 Predicting interested users

Once the comment section is separated from the main entry, it does not really need to be represented with the same vocabulary as the main entry, or for that matter, with the same type of the data (English words). Previously, we noted that there is much variety in writing styles in comments since they consist of multiple commenters. It is also quite possible that the difference in the commenter lineup is much influenced by the topic of the post. One way to express such notion in our model is to add user handles as another type of signal, encoding them as an additional set of observable variables which shares the topics:

$$p(\mathbf{w}_d, \mathbf{w}'_d, \mathbf{z}_d, \mathbf{z}'_d, \mathbf{u}_d, \theta_d) = p(\theta_d; \alpha) \cdot \prod_n^{N_d} p(w_{d,n}; \beta_{z_{d,n}}) \cdot p(z_{d,n} | \theta_d) \\ \cdot \prod_m^{M_d} p(w'_{d,m}; \beta'_{z'_{d,m}}) \cdot p(u_{d,m}; \gamma_{z'_{d,m}}) \cdot p(z'_{d,m} | \theta_d)$$

As before, this model postulates a set of latent topic variables, and each topic k has a corresponding multinomial distribution β_k over the vocabulary. In addition, the model generates the comment contents from a multinomial distribution β'_k , and a bag of users who respond to the post (represented as their user handles), from a distribution γ_k , both of them conditioned on the topic. In the experiment section below, we call this model **CommentLDA**. In this model, the topic distribution, θ , is all that determines the text content of the post, comments, and which users will respond to the post.

¹⁵ This also means that now the latent thematic concept, topic, is described by three different type of representation:

- 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.

The complete generative story of this model is the following: For each blog post d from 1 to D :

1. Choose a distribution θ_d over topics according to a symmetric Dirichlet distribution parameterized by α .
2. For n from 1 to N_d (the length of the post):
 - (a) Choose a topic $z_{d,n}$ according to θ_d .
 - (b) Choose a word $w_{d,n}$ according to the topic's post word distribution $\beta_{z_{d,n}}$.

¹⁵In another words, post text, comment text, and commenter distributions are all interdependent through the (latent) topic distribution θ .

3. For m from 1 to M_d (the length of the comments on the post, in words):
 - (a) Choose a topic $z'_{d,m}$.
 - (b) Choose an author $u_{d,m}$ from the topic's commenter distribution $\gamma_{z'_{d,m}}$.
 - (c) Choose a word $w'_{d,m}$ according to the topic's comment word distribution $\beta'_{z'_{d,m}}$.

On how to count users

In the above generative story, we designed the model so that a user handle is generated at each word position. The choice is rather arbitrary, and a few alternatives are possible. The model could generate a user handle once per post, allowing no duplicate of handle in the given document. It could also generate a handle once per each comment segment. Either method does not make much difference in inference complexity, but each assumes slightly different way of weighing the mode of participation. The first method (we call it here "count by verbosity") will force γ to give, other things being equal, higher weight to the users who write longer comments with more words (not just repeating one words). Other way of thinking this is that it will let garrulous users define the topics. The second method (we call it "count by response") is more democratic, letting every user who responds to the post an equal vote in determining what the post is about, regardless of the verbosity. The last one takes the middle road, giving more say to users who engage in the conversation repeatedly. (we call it "count by comments".) Each of those ideas will lead to a slightly different variation of the generative stories. They involved a change to step 3 in the generative process in the following way:

Counting by response (replaces step 3): For u from 1 to U_d (the number of users who respond to the post): (a) and (b) as before. (c) For ℓ from 1 to $L_{d,u}$ (the number of words in u_d 's comments), choose $w'_{d,u,\ell}$ according to the topic's comment word distribution $\beta'_{z'_{d,\ell}}$. This model collapses all comments by a user into a single bag of words on a single topic.¹⁶

Counting by comments (replaces step 3): For c from 1 to C_d (the number of comments on the post): (a) and (b) as before. (c) For ℓ from 1 to $L_{d,c}$ (the number of words in comment c_d), choose $w'_{d,c,\ell}$ according to the topic's comment word distribution $\beta'_{z'_{d,\ell}}$. Intuitively, each comment has a topic, a user, and a bag of words.

On (not) weighing comment contents

The above modification reflect the different weighing of the qualitative difference among the participants. Now, it is also possible to consider different weighing of the comment contents vis-a-vis user. If one believes that comment content is utterly uninformative in terms of understanding the thematic orientation of the post, it would make more sense to discount the count of comment contents. In fact, we could consider taking out this part entirely, resulting in the below model:

$$p(\mathbf{w}_d, \mathbf{w}'_d, \mathbf{z}_d, \mathbf{z}'_d, \mathbf{u}_d, \theta_d) = p(\theta_d; \alpha) \cdot \prod_n^{N_d} p(w_{d,n}; \beta_{z_{d,n}}) \cdot p(z_{d,n} | \theta_d) \\ \cdot \prod_m^{M_d} p(u_{d,m}; \gamma_{z'_{d,m}}) \cdot p(z'_{d,m} | \theta_d)$$

This does model which users are likely to respond to a post, but not what they will write. This model seems rather radical, however, it is actually much analogous to the model of [29]. (Though the variables are given much different meanings here.) In our experiment section, we call this model **LinkLDA**.¹⁷ The similar models were applied

¹⁶The 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.

¹⁷Instead of blog commenters, they modeled citations.

to different tasks in natural language processing research, such as relation extraction or polarity classification, with competitive results [89, ?]. In our application section, we will show that for some blogs, prediction results is better if comment contents are utterly discounted.

2.5.2 Results from user prediction

In [99], we applied the user handle prediction model to our blog corpus. In the following section we will give some highlights from the experiments.

Prediction performance

For each political blog in the corpus, we trained the three counting variations each of LinkLDA and CommentLDA. Model parameters β , γ , and (in CommentLDA) β' were learned by maximizing likelihood, with Gibbs sampling for posterior inference.¹⁸ The number of topics, K , was fixed at 15. We show the results for Matthew Yglesias (<http://www.matthewyglesias.theatlantic.com>, denoted “MY”) and RedState (<http://www.redstate.com>, denoted “RS”) in 2.2

As a baseline, we implemented a simple method which 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 strong baseline. We also compared to the results from Naïve Bayes classifiers (with word counts in the post’s main entry as features).

To perform the prediction task with the learned models, we first infer the posterior topic mixture for each new post (θ in the above equations), then computed each user’s comment prediction score as described in [99]. This let us order users by their posterior probabilities given the post’s words. We then used the top n users as our prediction. We report in Tab. 2.2 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.

In the table “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.

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 some of the site, including RS. LinkLDA usually works slightly better than CommentLDA, in all site except for MY, where CommentLDA is stronger. In RS CommentLDA is extremely poor. Differences in commenting style are likely to blame: MY has relatively long comments in comparison to RS (see the data size table in the previous section). 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

¹⁸For the detail of inference procedure, please consult with. [99]

	<i>n</i> =5	<i>n</i> =10	<i>n</i> =20	<i>n</i> =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
<i>Max</i>	<i>94.75</i>	<i>89.89</i>	<i>73.63</i>	<i>58.76</i>	<i>92.60</i>
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	11.30
Com-r	15.93	11.42	8.37	6.89	10.97
Com-c	17.57	12.46	8.85	7.34	12.14
<i>Max</i>	<i>80.77</i>	<i>62.98</i>	<i>40.95</i>	<i>29.03</i>	<i>91.86</i>

Table 2.2: Commenter prediction precision. The numbers are macro-averaged across posts. Each column contains results for the cut off value noted on the top. See the text body for more explanation.

than it helps CommentLDA. Varying the counting method can bring as much as 10% performance gain.

Descriptive aspects of the model

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 particular 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). We show words most strongly associated with a few topics, arguing that some coherent clusters have been discovered. Table 2.3 shows topics discovered in MY (using counting by comments). This is the blog site where our models most consistently outperformed the baseline, 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 2.3 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 [11].

The model clustered words into topics pertaining to religion and domestic policy (first and last topics in Table 2.3) quite reasonably. Some of the religion-related words make sense in light of current affairs.¹⁹ Some words in the comment section are slightly off-topic from the issue of religion, such as *dawkins*²⁰ or *wright*,²¹ 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

¹⁹Mitt 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. Muktada al-Sadr is an Iraqi theologian and political activist, and John Hagee is an influential televangelist.

²⁰Richard Dawkins is a well known evolutionary biologist who is a vocal critic of intelligent design.

²¹We 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.

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, financial, 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 2.3: The most probable words for some CommentLDA topics (MY).

in Table 2.3), 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 focus on the tangible (*oil, taxes, prices, weapons*).

2.5.3 Predicting post popularity

In the previous section, we designed the models which the generation comment contents and user handles. In this section we will develop models which capture aggregate attributes in comment activities.

Suppose one’s interest in a blog community is something simple, such as some *statistics* in word usage, or merely the aggregated *count* of the commenting user, rather than whom the post compelled to comment or what they write. In this task we will predict the volume of user comments, which in some sense is the simplest measure of post *popularity*.

²² As in the previous section we assume that the post content and the response values are connected through the shared set of topics. The key difference here is the distribution over the response values, which we postulate as a mixture of poisson distribution. Each component distribution is associated to a topic, therefore the document-level topic mixture, θ , also affects the mixture coefficient for the response generation. ²³ In the below generative story, the notations are the same as the previous section. We use m to represents the poisson mixture distribution. K refers to the number of topics as before.

1. For each blog post d from 1 to D :

- (a) Choose a distribution θ_d over topics according to a symmetric Dirichlet distribution parameterized by α .
- (b) For n from 1 to N_d (the length of the d th post):
 - i. Choose a topic $z_{d,n}$ from the distribution θ_d .
 - ii. Choose a word $w_{d,n}$ from $\beta_{z_{d,n}}$
- (c) For k from 1 to K , let

$$m_{d,k} \leftarrow \frac{\text{freq}(k; \mathbf{z}_d) + \alpha_k}{\sum_{k'=1}^K \text{freq}(k'; \mathbf{z}_d) + \alpha_{k'}} \quad (2.1)$$

Then choose a comment volume v_d from the mixture distribution $\sum_{k=1}^K m_{d,k} p(\cdot; \lambda_k)$.

Note that the model is identical to LDA until step c, where we define document-specific mixture coefficients and generate the volume from a mixture model over volume. This model is essentially a type of “supervised” or “annotated” LDA [12, 10, 65, 108], where side target variables are generated based on topics and therefore influence what topics are learned. The likelihood function of the above generative story is:

$$p(\mathbf{w}_d, \mathbf{z}_d, \theta_d, v_d) = \sum_{k=1}^K m_{d,k} p(\cdot; \lambda_k) \cdot p(\theta_d; \alpha) \cdot \prod_n^{N_d} p(w_{d,n}; \beta_{z_{d,n}}) \cdot p(z_{d,n} | \theta_d)$$

The inference of latent variables (and the model parameters) is slightly more complicated in this setting. In our application (see the next section), we have used stochastic EM with Gibbs sampling as before, while re-estimating the poisson parameters at each iteration and use them as the posterior distribution at the next iteration. See our publication [100] for more detail for the inference.

2.5.4 Results from popularity prediction

We applied our comment volume prediction model to the two sites from our blog corpus. Matthew Yglesias (<http://www.matthewyglesias.theatlantic.com>, denoted “MY”) and RedState (<http://www.redstate.com>, denoted “RS”). The full results were presented in [100]. In the following section we will give some highlights from this experiments.

Prediction performance

We use the training portion to estimate the model parameters. For a new (test) blog post, we infer its topic distribution θ , then compute the expected value for v , the comment volume, as described in [100]. For this application, we experimented with a few variation in comment volume. We report here the results with the comment volume measured by word token count and commenter count. ²⁴

²²[66] empirically tested the correspondence between this post popularity and the post comment volume.

²³Any distribution over integer, such as negative binomial distribution, can be used here.

²⁴More fine-grained predictions are possible as well (e.g., predicting the absolute number of words in the comments). Other future possibilities include the prediction of the rate of positive or negative polarity words, or words referring to a certain named entity.

		# words			# comments		
		prec.	rec.	F_1	prec.	rec.	F_1
MY	Naïve Bayes	72.5	41.7	52.9	42.6	38.8	40.6
	Regression	81.5	44.1	57.2	60.8	55.2	57.8
	T-Poisson	70.1 (± 1.8)	63.2 (± 2.5)	66.4	41.3 (± 2.1)	53.1 (± 3.5)	46.4
	k=30	71.8 (± 2.0)	60.1 (± 3.4)	65.4	45.3 (± 2.1)	54.2 (± 5.3)	49.3
	k=40	71.0 (± 1.9)	63.4 (± 2.7)	66.9	44.0 (± 2.1)	58.8 (± 3.3)	50.3
	T-NBin.	69.7 (± 2.3)	62.5 (± 2.5)	65.9	38.4 (± 2.2)	45.7 (± 3.3)	41.7
	C-LDA	70.2 (± 2.3)	68.8 (± 2.5)	69.4	37.2 (± 1.5)	50.4 (± 3.3)	42.8
RS	Naïve Bayes	64.1	25.7	36.6	37.8	34.1	35.0
	Regression	52.0	26.8	35.5	20.5	19.5	20.0
	T-Poisson	52.4 (± 2.8)	33.5 (± 2.0)	40.8	25.4 (± 2.6)	27.9 (± 2.9)	26.7

Table 2.4: Experiments: precision and recall for “high volume” posts. NB= Naïve Bayes classifier, Reg. = regression, T-Poisson = Topic-Poisson, T-NBin. = Topic-Negative Binomial, C-LDA = CommentLDA. Topic models are “ave. (\pm s.d.)” across 10 runs.

In this task we simply evaluate our model by how well it predicts whether a blog post will have “higher or lower” volume than the average. We note that our model is more powerful as it gives a distribution over values for v , permitting more fine-grained prediction and analysis. The mean volume is approximately 1424 words (35 comments) for MY and 819 words (29 comments) for RS. We report precision and recall for the “high volume” class. The results are shown in Table 2.4 for both blogs.

For this experiment, we have implemented several variations of the proposed model. In the table, “T-Poisson”, both $k=30$ and $k=40$, are our original prediction model described in the last section. “ k ” referred to the topic size, fixed at 30 and 40. “T-NBin” and “C-LDA” are the variations of this model. See the publication for more detail. (We will not recount the detail of those models in this proposal.) Our two contending models are naïve Bayes model and generalized linear model [32] (“naïve Bayes” and “regression” in the table).

On the MY data, our volume prediction model (“T-Poisson”) improves recall substantially over naïve-Bayes, on both measures, with a slight loss in precision. Its precision lags behind the regression model, gaining in recall on word volume prediction but not on comment volume prediction. The effect is similar on RS data when predicting *word* volume, but the loss in precision is much greater, and the model is ineffective for *comment* volume. Note that comment volume on RS is harder to predict from words. The regression model is much less effective on the RS data set, falling behind Naïve-Bayes on both tasks. The naïve Bayes model tends to err on the side of precision.

Descriptive aspects of the model

We contrasted the two blog sites, each from the opposite spectrum of American politics, in terms of issue popularities in each community. With our volume prediction model we can characterize each topic by λ_k (the mean for its Poisson distribution over volume values) and, more traditionally, by the words associated with each topic. Table 2.6 shows the topics discovered in MY by our model (using the word count as a proxy for the popularity volume). Topics are ranked by λ_k ; words are selected as in [11].

Table 2.6 also shows the binary accuracy on posts associated with each topic. We assign a post to each topic k that has $\theta_{d,k} \geq 0.25$ (a post can go to zero, one, or more topics), and measure binary prediction accuracy within the topic. These accuracies are based mostly on very small numbers of posts, so our analysis is tentative.

















λ_k	topic words	# posts	accuracy
1873	women black white men people liberal civil working woman rights	7	 (100)
1730	obama clinton campaign hillary barack president presidential rally senator democratic	13	 (77)
1643	think people policy really way just good political kind going	74	 (72)
1561	conservative party political democrats democratic republican republicans immigration gop right	12	 (50)
1521	people city school college photo creative states license good time	19	 (58)
1484	romney huckabee giuliani mitt mike rudy muslim church really republican	3	 (33)
1478	iran world nuclear israel united states foreign war international iranian	16	 (69)
1452	carbon oil trade emissions change climate energy human global world	6	 (33)
1425	obama clinton win campaign mccain hillary primary voters vote race	22	 (64)
1352	health economic plan care tax spending economy money people insurance	22	 (55)
1263	iraq war military government american iraq troops forces security years	24	 (58)
1246	administration bush congress torture law intelligence legal president cia government	5	 (20)
1215	mccain john bush president campaign policy know george press man	20	 (60)
1025	team game season defense good trade play player better best	8	 (38)
1007	book times news read article post blog know media good	23	 (43)
<i>Overall:</i>		183	 (58)

Table 2.5: MY Topic-Poisson model: Poisson parameter estimate and top words for each topic. See text for explanation.

The most comment-worthy topics on the liberal blog MY appear to deal with the gender/race issue and the Democratic presidential primary race. On the conservative RS blog, the Republican presidential primary race, and discussion about the party itself dominate. On both blogs, discussion of internal races for party nominations is clearly the most likely to incite readers of these partisan blogs to comment. Some clear issues arise as topics. On MY, the Middle East and energy are the seventh and eighth topics; healthcare is slightly below the overall average. RS rates religion very high (fourth topic), with the economy just above average and Iraq/Afghanistan well below. Note that the least comment-worthy topics on MY have to do with sports and reading material, which interest Matthew Yglesias but perhaps not his readers.

2.5.5 Summary of results

In this section we develop two sets of probabilistic prediction models for political blogs. The quantitative evaluations showed that those models are competitive to standard baselines. We also found that there are much difference in performance between the blog sites. Difference in blogging style, as well as the relative size of data are most likely the cause of those difference. The models which capture those differences during the learning process would perhaps be the next steps for this lines of prediction works. It is also possible that including different types of evidence, such as timing of the post, author identities, improve the model performance in either tasks. The latter idea may suggest much different parameterization from our current models, in order to better weigh among the different type of evidence.

Along with the quantitative evaluation, we reported here some of findings in the learned models. Those discoveries revealed much thought provoking patterns between the post and each community’s response. We built up our models from plain LDA, augment them so that the latent topics link the two halves of the blogging (post and comments). Because of this construction, our models can give illustrations on what topics would generate what reactions. In the case of user prediction task, the reaction is summarized by the set of users who would be interested in the

λ_k	topic words	# posts	accuracy
1546	romney huckabee mccain thompson rudy mitt fred campaign iowa mike	4	██████████ (75)
1378	mccain party conservative conservatives republican candidate john issues gop republicans	9	██████ (44)
1030	paul dan ron energy think oil thomas pick mclaughlin change	6	██████ (33)
977	man men america american life god great religion believe jesus	9	██████ (44)
954	court law justice supreme amendment general attorney state school states	8	██████ (50)
857	obama hillary clinton win democratic vote primary party race nomination	18	██████ (50)
846	economy market fed markets money rate mortgage prices inflation financial	15	██████ (60)
789	just way good people know right think time want say	146	██████████ (61)
692	tax health government care taxes state federal insurance spending money	14	██████ (50)
618	obama barack hillary campaign clinton senator mccain wright john church	44	██████████ (63)
594	just moe really going senator know lane right update man	35	██████████ (62)
592	iraq war troops qaeda iraqi intelligence afghanistan surge security general	7	██████ (42)
583	president trade united policy bush states foreign world israel iran	9	██████████ (77)
578	said hillary asked mccain host russert obama schieffer wallace barry	11	██████ (45)
545	democrats republican congress house republicans senate democrat rep gop year	19	██████ (36)
<i>Overall:</i>		231	██████████ (59)

Table 2.6: RS Topic-Poisson model: Poisson parameter estimate and top words for each topic. See text for explanation.

discussion of the post. We also show that the differences in blogger and commenter languages (or the different focus in the given issues) with commentLDA. In the case of the popularity prediction task, the reactions are represented as the size of comment volume; thereby we are able to make a probabilistic statement on what topics are interested in the community. We showed that the model can tell in succinct fashion not only which blog post would be popular but also why it become popular. Such site profiling technique is potentially useful for researchers who need to analyzing blog community or for content providers wish to keep track of the popularity trends among their readers.

We do not plan further experiments on either of our blog models in this dissertation; We hope future researchers in this area, as well as application engineers in the relevant fields, find our work here worthy contribution to the further developments.

2.6 Related works

In this section we note some of the most relevant works to our first task, blog comment prediction.

In topic modeling

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 [13] and has been especially popular because it can be understood as a generative model and because it discovers understandable topics in many scenarios [93]. 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 [106], opinion mining in user reviews [97], and sentiment discovery in free-text annotations [14]. [26] 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, [59] 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.²⁵ Several such studies are especially relevant to our work. [10] were one of the earliest results in this trend. The concept was developed into more general framework by [12]. [94] and [90] 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 [62] 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. [29], following [20], 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. [69] 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.

One of a common use of those models is prediction. Although unsupervised settings are much common among topic modeling researches, Some of them incorporate the observation on response variables during its training. The harbinger of the trend is aforementioned [10]. Mimno et al. ([65]) and Zhu et al. ([108]), also proposed LDA model with the labeled (or annotated) response variables in the generative story. Our volume prediction models [section reference] are similar to those model in overall approach (supervised training with latent topic variables), but differ in terms of what each variables represent, and how the response variables parameterized. For example, in supervised LDA [12] generalized linear model is incorporated in the generative story to represent the response variables, while in our model the response variables are sampled from a mixture of poisson distribution.

On politics and blogging

Network analysis, including citation analysis, has been applied to document collections on the Web [20]. [1] 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 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 [68, 60].

A few studies have focused on information in comments. The best known is perhaps Mishe and Glance [66]. showed the value of comments in characterizing the social repercussions of a post, including popularity and controversy. Their large-scale user study correlated popularity and comment activity.

Topic modeling has been used in political science research as a data exploratory tools quite often. Since they (predictably) discover “topics”, a set of distributions over the words associated to the thematic topic, as one of the consequence of parameter learning, they often present humanly interpretable, succinct description of statistical trends in the texts. For this reasons topic model has been used quite extensively for novel data exploration projects asides from prediction tasks. [81, 82, 67, 38]

²⁵Some of the models mentioned in the previous paragraph fit into this description as well

2.7 Conclusion and future works

In this chapter we presented the first part of our dissertation work, building of statistical model which predict the readership reaction from the original post in political blog sites. As of this proposal, the works described here were completed. The results were published in [99] and [100]. More complete version of the reports, including the technical details omitted from the publications, would be presented in full in the final dissertation.

The task we presented here have much practical as well as intellectual merits, in specific, with this task we achieved three contributions:

1. Delivery of a core analytical technologies which can be readily applied to such intelligent application as document prioritization of personal recommendation in and related to online media.
2. A new way of inquiry into the political blog forums and the political subcultures reflected on their written communications.
3. Application of latent topic modeling modeling to the type of social data which has much different characteristic from the conventional natural language research corpus.

Chapter 3

Proposed Works

In this chapter we present our proposed tasks.

3.1 Surviving Committee: Predicting Bill Success

Over the course of the legislative process, bills gather many response values from the various institutions (e.g., each chamber of Congress, the executive office). Among these institutions are the congressional committees. In this section, we propose the statistical models to predict whether an introduced bill will proceed further than the committee, or, in other words, whether the committee as a group will let the bill “survive”.

Here is the basic prediction setting: our predictor is given a congressional bill as introduced, and outputs the binary prediction on whether its referred committee will endorse it at some point within the Congress session (a two-year period). As was the case with the last task, our point here is to use the textual content of the document (the bill) to predict the repercussion (the committee action) it causes. Unlike the last task, however, we will use much more contextual evidence in this task along with the text contents.

The system of congressional committees constitutes one of the most intriguing systems in the legislative process. First of all, they are quite powerful. The rules of Congress grant them the great authority of assessing whether bills deserves to be *considered*, and therefore be voted on publicly in the chamber as a whole. [16] Secondly, the decision-making mechanism inside the committee is rather mysterious. Unlike the floor voting, there are no clear-cut procedures for committee members to reject bills. Also, it is difficult to track their internal conduct records in large scale. The main evidence you can observe in this process is that the bills went to committee, and whether or not the bills came out. Since a bill is mostly made up with texts, we believe that inquiring into how text analysis makes a difference in this task is quite worthwhile.

In the following sections, we will discuss congressional committee functions, insights from political science, and how we propose to capture the relationships between the bills and committee responses in our prediction models. Since this part of the work is not complete, we will describe only the basic models at this time. We will discuss possible alternatives and extensions to the proposed work in the last section.

3.1.1 Background

Every Congress, the House of Representatives discusses somewhere between 500 to 1,000 bills on the floor. Once scheduled for the floor debate, a bill’s course of life is somewhat prescribed, if not predictable. Upon the motion for a recorded vote, representatives cast their votes (“Yea or “Nay”) for the passing of the bill. If the majority agrees to pass

the bill as it is presented, the exact version of the bill is sent to the upper chamber, Senate. If the Senate also agrees to pass this version of the bill following the same procedure, the bill will become a public law upon the approval of the executive branch. Whenever rewrites (revisions, amendments) are necessary, the new version of the bill repeats the rounds of voting until all parties agree upon a final version. [16]

This string of political action is, by and large, a public event. A trace of recorded votes (“roll calls”) from the current Congress, as well as historical Congresses, are kept by officials and released to the public via the Library of Congress. Concerned citizens pay close attention to the roll calls since they directly affect the course of the nation and the welfare of many. Quantitative political scientists study extensively patterns in roll call voting, and examine their mechanism through the model building. Roll call records are considered yardsticks to measure different politicians’ policy and issue positions. The record keeping also offer one way of holding members of Congress accountable, as they are praised and criticized for the votes they made. Often, past votes are used as political weapons, for they can indicate contradictions, flip-flopping, compromises, or a lack of integrity.

All in all, those roll call records are important resources for studying the legislative process. However, they do not represent the whole spectrum of the issues. In fact, roll call records reflect only a fraction of the bills that are introduced. What happens to those other bills? Most of them die at the **congressional committees**.

Upon introduction to the legislative system, each bill is assigned (“referred”) to one of the congressional committees for examination. The referred committee, in turn, deliberates on the merits of the bill, and then reports back to Congress if the committee decides the bill deserves further consideration. The legislative rules also give the committees the power to rewrite bills as they see necessary. Some 4,000 to 8,000 bills are introduced to the House of Representatives each Congress. On average, 85 percent of those bills effectively die with the referred committee, receiving no further attention.¹ Interestingly, once a bill hits the floor after surviving the committee, it has nearly a 90-percent chance of eventually pulling through the chamber, making the committee the toughest hurdle in a bills life.

Because of this extensive pre-screening responsibility, committee membership, especially the chairmanship, gives tremendous power to a politician. In particular, the committee system is often described as **agenda setting** system [2, 51], since it could prevent certain issues from ever going into public discussion, hence effectively branding them as non-issues.

Despite the committees’ significant authority, the working inside the committees is rather opaque. Unlike the floor voting, where the rules clearly stipulate what actions are necessary to “pass the bill” (a majority vote of the Congress members present), the rules governing the committee decision-making process is much more elusive. Moreover, politician conduct inside the committees is more difficult to track, and therefore, subject to much less public scrutiny. Recorded votes in committees, if such happens, are unevenly kept, and historic records on the discussions are not standardized, making it harder to compare across the committee and congresses.

Naturally, there is much discussion among political scientists on what really is the force behind committee politics [2, 51, 41, 30]. What are the working principles of the committees? What are the systematic constraints and what are the individual variables? Do the issues in the bill drive the committees decision, or is that decision solely a function of personal relations among the actors? And, most of all, how do those various factors relate to the final outcome by the committee as a whole?

Throughout the course of our model design, we will keep in mind *how* the model be useful in answering those questions.

¹“Effectively dies, but in reality, no bills are ever officially killed. They are simply left expired at the end of that session of Congress. These bills could also be resurrected in a later Congress under another guise.

Congress	Majority	Introduced (R) (D)	Pass (R) (D)	Rate (R) (D)
103	D (Clinton)	5,311 (1,856) (3,455)	622 (64) (558)	0.117 (0.034) (0.162)
104	R (Clinton)	4,345 (2,426) (1,919)	596 (479) (117)	0.137 (0.197) (0.061)
105	R (Clinton)	4,875 (2,796) (2,079)	643 (531) (112)	0.132 (0.190) (0.054)
106	R (Clinton)	5,682 (3,299) (2,383)	859 (691) (168)	0.151 (0.209) (0.070)
107	R (Bush)	5,768 (3,104) (2,664)	697 (542) (155)	0.121 (0.175) (0.058)
108	R (Bush)	5,432 (2,915) (2,517)	759 (611) (148)	0.140 (0.210) (0.059)
109	R (Bush)	6,437 (3,652) (2,785)	759 (618) (141)	0.118 (0.169) (0.051)
110	D (Bush)	7,341 (2,668) (4,673)	1,065 (226) (839)	0.145 (0.085) (0.180)
111	D (Obama)	6,571 (1,949) (4,622)	828 (158) (670)	0.126 (0.081) (0.145)
Total	-	51,762 (24,665) (2,7097)	6,828 (3,920) (2,908)	0.132 (0.159) (0.107)

Table 3.1: Count of introduced bills per congress, along with the committee survival rate, and break down by the bill sponsor’s party affiliation.

Motivation

We believe that building prediction models for committee actions makes good sense in two ways. First, it can be quite useful in practical manners to identify the bills which will most likely to make it through at the early stage, to explain reasons for bill survival, and to do so without expert knowledge or the lengthy perusal of thousands of bills. After all, once a bill hits the floor, it is already 90 percent of the way to the finish line. If one wishes to effectively monitor the rise of issues and changes in political climates, identifying the bills that matter before they become apparent might be desirable.

In the previous chapter, we examined the various model parameters after fitting them to the training data. We gained valuable insights into the response generative process through those explorations. We intend our bill prediction model to be useful in this respect as well. This purpose, gaining insights into the mechanism in the decision process, is in fact the primary motivation for the many predictive statistical models for roll call voting, most notably the ideal points estimates such as NOMINATE [78, 77, 21] or Bayesian NOMINATE [45, 18]. Compared to floor voting, committee level decisions are hard to analyze in statistical means due to the lack of descriptive evidence roll call records provide. In this task, we proposed to bridge this difficulty by utilizing textual evidences.

3.1.2 Congressional Committee Data

To support our data driven research on committee survival prediction, we have collected all the bills introduced in the House of Representatives from 103rd (1992) to 111th congress (2010).²

We have gathered all version of all the bills that ever appeared in the chamber during this time period. For our current experiments we use only the introduced version of bills (titles and contents).³ All our bills were downloaded directly from the Library of Congress’s Thomas website (<http://thomas.loc.gov/home/thomas.php>). Along with the bill title and the text, we also gathered the legislative histories (including roll call and other major actions) for each bills.

We obtained our information about the committee between 103rd to 111th from Charles Stewart’s resource at MIT. (http://web.mit.edu/17.251/www/data_page.html).⁴ Additional information are supplied by John

²Several congressional bill corpora exist, though we have not found a corpus that includes both the bill text with other meta data.

³Over the course of its life, bill contents can change significantly. Even the title of the bill also changes sometimes.

⁴The House Clark office is in charge of those records, though they do not maintain publicly accessible online resource.

Wilkerson' Congressional Bill Project (<http://congressionalbills.org/>) at University of Washington.

In our corpus each bill is matched up with its texts, title, committee assignments, and a binary value indicating whether the bill is referred back to the full Congress (i.e. "recommended") from the committee or not. We also extracted several meta contextual data, such as sponsor's name or party affiliation, from each bill's summary page provided by the Library of Congress.

There were a total of 51,762 bills introduced in the House during this period. A total of 6,828 of them have passed the committee and progressed further in the legislative process. See table 3.1.2 for the break down by congress and party. The average rate for a bill passing committee is 13.2%. The best passing rate was the 110th congress during the George W. Bush administration in 2006 to 2007, with 14.5% of bills passing through committee. The party membership breakdown of each committee corresponds to the party division of the chamber as a whole. Committee chairs are selected by the majority party caucus. Membership selection is left to each party's discretion.

Preliminary data analysis

Policy specialists often cite the importance of the personal skills and personal connections of politicians in discussing the success (or failure) of a bill. As a sanity check, we have examined whether simple personal attributes of (main) sponsor separate the winning bills from the losing bills. The most obvious attribute is perhaps party affiliation. Our data shows that sponsor's party affiliation is a good indicator for the increased chance of survival, though far from decisive. This indicates that committee decisions were not operated by strict party control. In fact, our data shows that there are more failed bills sponsored by majority party than minority party:

	Sponsor in majority	Sponsor in minority
Survive	5,543 (10.7)	1,285 (2.5)
Die	25,435 (49.1)	19,499 (37.7)

Table 3.2: Total number of survived bills, by party affiliations. Numbers in the parenthesis are the percentage in all the data

What if the main sponsors are himself in the assigned committee? We found that many successful bills are sponsored by a committee member, but it too does not clearly divide the survivors from the casualty:

	Sponsor in the committee	Sponsor not in the committee
Survive	4,247 (8.2)	2,581 (5.0)
Die	16,916 (32.7)	28,018 (54.1)

Table 3.3: Bill survival rate by sponsors committee affiliation.

We also checked if party identity (membership in either the Republican party or the Democrat party) is a good indicator, but the difference between Republican party membership and the Democrat party membership is much smaller than majority affiliation or committee affiliation. The other personal attributes, as well as some non-personal attributes, we have checked in this manner are seen in table ???. None of these attributes clearly separate the data, though some of them appear more with the successful bills than with unsuccessful bills. Clearly, there are some mannerisms in the committee decision making. However discovering the finer patterns in these mannerism seems to require more than perfunctory investigation. We will investigate how the statistical model based on texts analysis improves the prediction.

3.1.3 Task

As a part of this dissertation work we propose the development of statistical models for predicting congressional committee reactions to a proposed piece of legislation. We assume the model performs its task in a strictly predictive settings: The models are to output a prediction about the referral committee’s reaction (e.g., “recommended” or “not”) given a previously unseen bill. The bill information available at prediction time includes the words in the bill title and texts (as introduced), the name and party of the main sponsor, name of the referred committee, members of the committee, and other contextual meta data presumably available at the time of the legislation’s introduction. The models are not given any information on the legislative actions on the bill beyond the referral assignment. The models may use historical data on legislative actions along with textual and contextual data, for the purpose of training.

We evaluate the model by the accuracy, precision, and recall on how many of the predicted bills actually survived the committee. We use data from the 103rd to 110th United States Congress as training data, and 111th Congress (the last Congress in the data set, from 2009 to 2010) as evaluation data.

3.1.4 Proposed works

We believe that the contents of a bill have much to do with its survival though committee. This is a quite intuitive notion since, after all, a bill is mostly text. However, this is not a perspective often seen in the studies about bill success. Many of them focus on politician’s personal connections or institutional powers in the legislative process.[2] We agree that these are important factors. However, we believe that there is a great deal of untapped power in the textual information of the bill itself.

In this work we will develop three sets of experiments to investigate the utility of textual contents to predict the bill’s legislative actions at the assigned committees:

- Generative text modeling approach (an application of the volume prediction from the last task)
- Text regression with a log-linear model
- Latent variable log-linear model

We believe that these sets of investigations will cover a good ground in examining predictive powers in the text. We will discuss each of them in turn in the following sections.

Generative text modeling approach

The first approach we will examine in this task is generative text modeling. This is essentially the approach we used in the first task. In the blog comment prediction tasks we cast the task as an instance of topic modeling, and build the prediction model as a probabilistic inference on this model. In this scheme the prediction was made based solely on the words in the given post.

In fact it is quite possible to use this same approach to the current prediction task. Though the legal documents do not look like blog post, the processes we want to capture are actually quite similar: generation of responses toward a piece of writing by the group of people who presumably read the writing. Our comment volume prediction model can be applied here with small adjustments. The only change in the model is the semantics of the variables, but the structure of the model remains the same in the new task. Bill contents will replace the post contents, and the response values will become binary indicator of whether the bill survived or not. We will need to replace the Poisson random variable with a more appropriate distribution, and the inference steps will change accordingly. We may need to employ new optimization techniques. Overall, however, we estimate the change required for this application to be quite manageable.

We have two reasons to believe that this can be an effective approach, hence worthwhile to cover in these experiments. First, we have the first-hand knowledge of using this approach. In our experiment we showed that prediction with this model performs quite competitively. We also demonstrated that the fitted latent variables provide humanly understandable explanations of the response process. Second, recently a topic modeling approach was applied successfully for the prediction of legislative actions. Gerrish and Blei used a variation of supervised LDA in their roll call (floor) vote prediction tasks [34], and achieved competitive results. Their model used exactly the same type of textual data as our task, though their prediction target is a different stage of legislative action. For those reasons we believe generative model is a natural place to start for this task.

Text regression with log-linear model

The second approach we will experiment in this task is text regression with log-linear models.

Our preliminary examination of committee data showed that personal attributes (such as party affiliation or committee membership), though far from decisive, are good indicators for bill survival. There are many more such possibly indicative factors. For example, It is a commonly held suspicion among the policy specialist that the committee decisions are much influenced by the politician’s personal friendships or “ties” [19, 107, 15]. Other studies suggest that in some committees, regardless of their members, it is just much easier to pass a bill than others [2].⁵ The Committee on Appropriation, for example, passed nearly 50 percent of its assigned bills. There are possibly many other such clues. This suggests two desirable criteria for our prediction model:

- The model needs to be able to incorporate arbitrarily many heterogeneous types of evidence
- The model needs to be able to weigh among these evidences.

Generative approaches like the one from the last section are great for modeling the process, but are not particularly flexible in dealing with heterogeneous evidences. Even though they *can* incorporate different types of variables (like we did in the user prediction or volume prediction), each new addition requires an additional random distribution, entailing cumbersome reconfiguration or redesigning of the model structure. Instead of such compound distributions, what we want here is *one* flexible distribution which can represent many types of evidence in one random distributions.

Log-linear parameterization is used often in statistical natural language processing research for the problems with similar requirements.⁶ This is what we take on as the second experiment.

Previously with the generative framework, the conditional distribution over the prediction target was *derived* via probabilistic inference. In this experiment we directly model the conditional distribution over the response variable with a log-linear distribution.⁷ Let y be our target response variable, and \mathbf{x} be the document contents, and $\mathbf{g}(y, \mathbf{x})$ be a set of indicator features. Then, in the most general terms, the conditional log-linear model is expressed as:

$$p(y|\mathbf{x}) = \frac{\exp(\mathbf{w} \cdot \mathbf{g}(y, \mathbf{x}))}{\sum_y \exp(\mathbf{w} \cdot \mathbf{g}(y, \mathbf{x}))}$$

At test time, prediction is computed simply by finding the argmax over y . Since the current task is a binary prediction task this part is easily solvable.

$$\hat{y} = \arg \max_y \log p(y|\mathbf{x}) = \arg \max_y \mathbf{w} \cdot \mathbf{g}(y, \mathbf{x})$$

⁵We confirm with our data that there are quite a wide differences between committees in terms of its passing rates.

⁶Linear regression and logistic regression are closely related learning algorithms. In fact, Log-linear models can be seen as a probabilistic interpretation of those models with a log-loss function.

⁷The fundamental difference between this model and the previous generative model is that a conditional model is agnostic to the generative explanation on the observable (input) data, while in generative framework the distributions over the inputs are explicit. This entails much difference in model design, model limitations, and inference procedure.

contextual features	if the bill sponsor is affiliated by Republican party
	if the sponsor is the same party as committee chair
	if the sponsor is a member of referral committee.
	if the sponsor is congressman i ($\forall i \in \text{all member House}$)
	if the sponsor is from the state i ($\forall i \in \text{US state}$)
	if the bill is introduced in the month i ($\forall i \in [1..12]$)
	if the bill is introduced in the i year of the congress ($\forall i \in \text{1st, 2nd}$)
	if the bill is referred to the committee i ($\forall i \in \text{all House Committees}$)
	if the bill is secondary referred to the committee i ($\forall i \in \text{all House Committees}$)
content features	if word i is in the body of bill text, ($\forall i \in \text{Bill vocabulary}$)
	if word i is in the title of the bill, ($\forall i \in \text{Title vocabulary}$)

Table 3.4: Feature functions used for the preliminary experiments.

There are many way to train this model. A simple method is to form a log-loss function, and optimize it with respect to the model parameters [91]. Even though in general these loss functions do not have a closed form solution, efficient iterative algorithms are available. For the moment we do not discuss the inference.

Preliminary experiments with standard log-linear model

The important decisions in log-linear modeling is the design of feature functions, $g(\mathbf{x}, y)$. As we will see later in this chapter, particulars in the feature functions can reflect a variety of assumptions in the model. For the first round of the experiment, we designed and implemented a simple, basic set of features based on the sponsor information (e.g., sponsor affiliations) or institutional (assigned committee, assignment timing) information on the bill. We call them *contextual* features. All feature functions are indicators function, for example:

$$g_i(\mathbf{x}, y) = \mathbf{1}(x_{\text{sponsor}} = \text{republican}, y = \text{survive})$$

In addition, we use all the words in the title and all the words in the bill texts and titles. We call them *content* features.⁸ See table 3.4 for the complete list.

As seen in the table ??, bill survival has a very skewed distribution. A simple model which always outputs negative prediction can trivially achieve 86% accuracy. (In our tests set, 111th Congress, fixed negative prediction achieve 87.3% of accuracy.) As noted in section 3.1.3, we use 103rd to 110th congress as training data, then we evaluate the model by how well the trained model predicts the bills in the 111th congress. We show some of our results in table 3.5.

Type of features	Accuracy	F1 score	False Pos.	False Neg.
Context + Content	89.9	47.7	131	530
Context feature only	87.4	1.4	3	822
Baseline (fixed prediction: die)	87.3	n/a	0	828

Table 3.5: Accuracy and F1 of preliminary experiments.

We have manage a small gain over the fixed-negative prediction (baseline) with combination of content features and context features. We achieved the best prediction accuracy (89.9%) with the combination of all the text and contextual features. Using context alone does not do as well as using the text feature alone, although seems to add much discriminative power when combined with the text.

⁸we standardized the text by removing some stop words and rare words from the vocabulary, and convert all the words into lower case.

Our experiments suggest that both content and context are at work in the committee decision process. We are currently planning to expand the feature sets to include more sophisticated functions informed by the existing discussions in policy research. We expect such features as cosponsorship ties, seniority ranking, or regional ties between the sponsor and committee members [5, 107, 2, 61], could make a difference. We also plan to enlarge the content features with more intelligent text processing and standardization. Comparison between the two flavors of features, context and content, could bring out new insights on the committee decision process.

Latent variables log-linear model

One of the insights we learned from policy researchers is that certain types of bills are much easier to pass by virtue of their function or their objective, while others have a difficult time without substantial consensus building. The most apparent examples of “easy” bills are such bills as the “naming” of “post office” or the “honoring” some historical events or individual. These bills tend to advance easily since the committee has no political reason to reject them. Some cyclical bills, such as budget or appropriation, also tend to advance but for a different reason. Many of those bills are time-sensitive, or critical to the function of the government, therefore tend to be well facilitated.

In [2], Adler and Wilkerson discuss different types of bills and how the type of a bill makes a difference in the bill’s success or failure. They suggest three main bill categories (“discretionary”, “trivial”, “must act”) in terms of urgency. Others suggested categorization along the different types of quality, such as importance or scope.[80, 42, 71]

These studies suggest that committee decision mechanisms may depend on some latent, functional categorization of the bill. Thus, for the last experiment, we propose committee survival prediction models based on these insights. We will extend the conditional log-linear model from the last section with a set of latent random variables.

The merit of this approach is threefold. First, latent variables are quite a natural way to encode this insight insight into the model. Second, this idea, using “auxiliary” labeling or categorization to help target tasks, is actually a well studied technique in natural language research, and sentiment analysis in particular [73, 105]. Pipelining the two architectures, and training them separately, has been used often [105]. More recently, joint trainings are used successfully [104, 64]. Some of these approaches also involve latent variables [102]. Use of latent variables with conditional log-linear models is less common than with generative models, though there are some notable natural language works that successfully employed the tactics [76, 23, 27]. We expect our work would benefit from those studies. The third reason is that latent variable log-linear model would make a nice complement to the other two experiments. Latent topic model is more expressive, however, we trade off this expressiveness for the flexibility for log-linear model at the second experiment. Latent log-linear models may turn out to be the best of the two.

Encoding insights with feature function and latent variables

Let y be the response variable in question (in our case, bill passage). Let’s for now assume a classifier which predicts the bills’ functional categorization. Let z be an indicator variable of this category, and \mathbf{x}_{bill} be the content of the document. Then, what we would like to figure out first is $p(z|\mathbf{x}_{bill})$. To be sure there is no technical constraint says that the category has to be conditioned on the bill’s contents. But for the moment we assume this relationships between bill content and category.

One of the debates over the committee politics is whether it is the issue (or bill contents) or the personal skills (say, person’s connection) that matters the most in survival. We are not experts on the issue, but in designing the current model we assume that, once the bill’s functional type is decided, the only things that can change the course of the bill is personal connections. Lets call those attribute \mathbf{x}_{ppl} . Then a stochastic model over the bill’s survival is simply expressed as:

$$p(y|\mathbf{x}) = \sum_z p(y, z|\mathbf{x}) = \sum_z p(z|\mathbf{x}_{bill}) \cdot p(y|z, \mathbf{x}_{ppl})$$

Expressing this in log-linear parameterization, we will have something of this form:

$$p(y|\mathbf{x}) = \frac{\sum_z \exp(\mathbf{w} \cdot g(z, \mathbf{x}_{bill}) + \mathbf{w}' \cdot g'(z, y, \mathbf{x}_{ppl}))}{\sum_y \sum_z \exp(\mathbf{w} \cdot g(z, \mathbf{x}_{bill}) + \mathbf{w}' \cdot g'(z, y, \mathbf{x}_{ppl}))}$$

The second one is actually more expressive than the first one. Any parameter values expressed with the first one could find a equivalent in the second, but the reverse relationship does not hold. Note that this is just a very standard log-linear model with latent variables (which is even more expressive):

$$p(y|\mathbf{x}) = \frac{\sum_z \exp(\mathbf{w} \cdot g(y, z, \mathbf{x}))}{\sum_y \sum_z \exp(\mathbf{w} \cdot g(y, z, \mathbf{x}))}$$

with just a little more specification on feature function design. The *specifics* of the design reflect what our particular assumptions are in the response generation mechanism.

In the above we assumed that the bill’s contents are solely responsible in deciding the bill category. We can, for example, relax this assumption, allowing the bill content to influence the bill survival, not just the categorization of the bill. Another possibility is to switch the sets of features, and have contents influence the committee decision, while contextual features influence only the selection of bill category. Those change each represents slightly altered assumptions in committee decision process. In our experiments, we will explore these alternative assumptions and compare their performance. Those comparison, we hope, could bring out new insights on the committee decision process along with the better prediction performance.

3.1.5 Conclusion and future work

In this chapter we presented the first proposed task in our dissertation, the building of statistical model which predicts the committee decision based on the introduced version of a congressional bill. We communicated our motivation, our approach, and our outline of the proposed model which we will deliver in the final work. Upon completion of this task we aim to achieve the following.

1. Design of a core analytical method which can be readily applied for data-driven inquiries in the committee decision making process.
2. Introduction of novel text-driven prediction task with high social impact to the natural language community.
3. Development of flexible social prediction framework based on a latent log-linear model.

In this proposal we have reported some of the completed works as well as the development sketch of our planned works. At the time of this proposal, data collection, preliminary data analysis, basic model designs were completed. Implementation of the model, model evaluation, model refinement, and model extension are currently in progress.

We will first implement and evaluate the model we have discussed. We will also experiment with several variations of the hypotheses we have laid out in the section. In our framework, different hypotheses on the committee decision making can be explored in a principled manner by augmenting the feature set, by redesigning the feature function, and constraining its linkage to the latent variables. We hope to see much other types of insights beyond those presented in the last section as an extension of our work. We plan to publish our experimental results in computer science conferences, followed by either comprehensive or supplemental material along with the final dissertation. We hope to release our code base and data around the same time.

3.2 Motivated Voters: Predicting Polling Decision

In this section, we propose prediction models for a change in a person’s voting preference during a political campaign, and how the media exposure is related to this change. The aim of these models is to predict whether a voter will regard his original candidate choice favorably or unfavorably at the later polls. In other words, whether the voter will “flip” his favor or not.

Needless to say, the political election is one of the most exciting subjects in the prediction industry. The advent of web 2.0, and sentiment research, has turned this old pastime into a new challenge in natural language processing research. Can we predict the outcome of an election (or the public poll results) from the contemporaneous online contents? Similar question are asked in the social sciences, and political psychology, but with much of the emphasis on the process of change, or *How* people react to the information in the media.

Voting prediction is certainly one of the most fascinating types of response prediction; however, it is much more difficult to implement with the use of naturally available data only. Usually, the causation between the online media coverage and the general poll results is too indirect or ambiguous, making it difficult to represent the relation as a probabilistic system. In this work, we will use controlled user experiment data, thereby avoiding this problem of ambiguity. We will use this data also for model evaluation. In comparison to our other two tasks, we will pay much closer attention to the social science theories in this task. Specifically, the basic design of our probabilistic models is largely influenced by ideas from theory, and the user experiments which provided the data.

In the following sections, we will first discuss the theory of motivated reasoning in the context of political media campaigns. This is the political psychology theory we will work with in this task. We then will outline our prediction task. Discussion on the model building and experimental process will follow. Then, we will outline our proposed work. As before, we will summarize our development plans and several possible extensions in the last section. The work we present here is a part of our interdisciplinary project with political behavioral psychologist David Redlawsk and his team at Rutgers University. The project is a part of Social-Computational System Program at National Science Foundation.

3.2.1 Background

How do people decide who to vote for? There are literally thousands of opinions on this subject. Perhaps one of the most widely held beliefs is that media campaigns matter. Judging from the exorbitant amount of resources spent each year on political campaigns, this must be the case among the politicians. But how does a campaign really work on peoples’ minds? How, if ever, do campaigns persuade people to vote differently at election time?

During political campaigns, voters are exposed to a massive inflow of candidate information. Given a set of issues and personal preferences, a rational adult would examine the information carefully, and cast the vote to the candidate who presumably best serves their interests. The process may be disrupted by a lack of information or difficulty in comparison; nonetheless, the goal of the rational voters’s minds should always be to find the option which maximizes the utility with respect to their interests.

There are, however, indications that this “rational” model is not quite the whole story in human decision-making. A number of political psychology researches argue that, in the face of affectively incongruent information (negative information on something one already favors), otherwise rational human beings often behave irrationally, treating evidence with markedly impaired reasoning [85, 86].

It is often suggested that negative information about the favorites, when taken in a small dosages, may indeed *strengthen* the voter’s support to the candidate [86]. This human phenomena is often discussed in relation to more general social psychology tendencies such as **motivated reasoning** or confirmation bias [53, 63]. Another theory

of the political decision process, affective intelligence, suggests that voters tend to be more alert when challenged with incongruous information. They behave, paradoxically, much like attentive information seekers under anxiety, and chose the alternative which is closer to the ideal rational choice. Although fundamentally in agreement with the affective intelligence theory, the students of motivated reasoning caution that the anxiety-boosted alertness does not initially lead to better decision making; the state of affective intelligence comes into play only after a certain **tipping point**, the threshold of incongruence, is reached.

In recent years, a group of political psychologists initiated a data-driven inquiry into this subject: if we know *all* the articles a voter read during a political campaign, know all of his original issue positions and other relevant information, and know his decision at the polls, can we then discern the tipping point effect and motivated reasoning in the voter's behavioral trace? The Dynamic Process Tracing Environment (DPTE) project [87] was created to answer such a question. The resource is unusual in a research field where various types of surveys are the regular data collection methods. Unlike those methods, which rely on self-reports from research subjects, DPTE data aims to indirectly gather data by observing subjects' browsing behavior during a simulated political campaign.

The studies based on the data gathered with DPTE shows that motivated reasoning's effect and its relation to the contents of what the subjects read can be detected [86]. These results also suggest that latent cognitive states (and the changes between those states) can be detected from some observable behavioral trends, such as browsing histories. Do those findings mean we can predict how an individual is going to cast votes from a statistical analysis of their browsing trace? Or we can tell when they *will* change their minds? What can we find out about the voter's decision-making process from such prediction model? In this task we aim to answer these questions by building such models.

Motivation

Voting prediction model is in its own right an interesting pursuit; however, we would like to add two more motivations in undertaking our current task.

In the previous chapters we studied two distinct types of probabilistic model, conditional model (log-linear model) and generative model (topic model), for the purpose of social response predictions. We showed that the two techniques are versatile, and together they could conceivably cover a broad range of prediction problems. One common weakness in both approaches, however, is their implicit assumption on data independence. The assumption, though convenient in many situations, makes it difficult to capture temporal effects in a stochastic process. In motivated reasoning theory, and in our data based on DPTE experiments, it is much more natural to view the voting outcomes (per individual) as a sequence of actions, not a set of isolated instances. In our experiments we will study how this alternative data representation fare in response prediction tasks. We expect that the exploration into this direction increase general usefulness of our work in the broader research context.

In this task we mainly describe a statistical model to predict a specific voter behavior mapped to a single binary indicator, namely, whether an individual will vote for the same person he or she initially chose. This is certainly a quite useful variable to forecast. However this is not the only thing with which students of motivated reasoning, or other voter behavior theories, concern. In our DPTE project in particular, we have recorded many more other behavioral, as well as emotional, traits than a binary voting decision. (See the section 3.2.2 for overview. The through discussion of the user studies is in [86].) Our long term plan in this project is to extend our investigation further in the future to include these additional observations. We intend our current study of prediction modeling to shape the base of many others we will conduct in the future. We will discuss examples of possible extensions further later in this section.

Issue position	Economics and trade
	Defense and foreign policy
	Healthcare and other entitlement
	Other domestic issues
Commentary	on campaign performance
	on personality
	on debate performance
Others	Biographical Information
	Campaign slogans and motto
	Media polling reports and approval rate
	Party affiliation (D/R/I)
	Endorsement (organizational)
	Picture (portrait of the candidate)

Table 3.6: Types of articles used in the DPTE simulation studies.

Healthcare & entitlement	Health policy, Welfare reform, Drug policy, Poverty issue, Homeless issue, Entitlement programs
Biographical information	Age, Religion, Families, Political experience, Education, Experience prior to politics, Military experience, Ethnicity

Table 3.7: Examples of the articles subject.

3.2.2 DPTE: Dynamic Process Tracing Environment

To support experiments concerning the effect of motivated reasoning and other voting behavioral theories, a series of user studies have been conducted using the Dynamic Process Tracing Environment (DPTE) proposed in [87].

During DPTE experiments, subjects were presented with a set of mock political campaign simulations. For 25 minutes of a fictional “primary”, the subjects (“voters”) were instructed to read from the random flow of articles to decide which candidate to vote for at the end. At a fixed interval they were also asked to take polls, to indicate who was their current favorite contestant. The current system includes four polls. The last one is presented as the final primary vote.

During the primary, each subject is placed at a computer terminal, on which article headlines continuously scroll down. The subject may click any of the headlines to view the detail of the article while it is on the screen. The headlines say what the article is about, but does not reveal the detail of the articles. For example, a headline may relate that the article is about “candidate Singers political view on economics. The subject needs to click the article headline to read the article. The system is designed so that the subject experiences natural information seeking and browsing activity during the real political campaign. For each experiment, practitioners can adjust the article contents and other settings to suit for their later analysis.

The set of data we will use for this proposed work is composed from the DPTE experiments described in [86]. During these experiments, unbeknownst to the subjects, certain portions of the simulation articles were manipulated to relate subject-specific incongruent information. For example, a manipulated article could reveal that the candidate to whom the subject had originally expressed her affinity is a supporter of some unfavorable cause. The purpose of this manipulation was to ensure the subjects be systematically exposed to certain amount of incongruous information. The

amount of the manipulation were varied from zero up to 80%.⁹

207 non-student subjects behavioral traces (such as the type of articles read, level of manipulation, time spent on each article, polling results), as well as the subjects profiles (issue concern, self-claimed personal traits, party affiliation, demographic identification) were collected during the experiments. In addition, we had the subjects to give more detailed account of their voting preference (such as a ranked preference of the candidates). On average, the subject read 93 articles during the course of experiments. The article includes variety of topics. (See table 3.6 and table 3.7). We will discuss more about the representation of the data later in the chapter. See [87] for more through discussion of this user experiments.

From these data, we use the subjects' browsing trace, polling decisions, and some profile information to create a set of paired examples. See the description in the preliminary experiments later in the section.

3.2.3 Task

As a part of this dissertation work we propose the development of statistical models for predicting an individuals voting preference change during a political campaign. Specifically, we will ask if voters will defect from their originally favored candidate as response to what they read in the media.

The models are to output a sequence of predictions on an individual voter's "decision" given the profile of the voter and a score of articles he has read. Each decision is represented as a binary value indicating if he favored the originally liked candidate at a pole. The articles are given as a sequence as well. The models are to know when the decision occurred in relation to articles. The models may treat each sequence as one output, or each piece of decision as one output. At the evaluation time, performance metrics are based on the count of correctly predicted piece of decision.

The information on the voter includes who he initially liked and disliked. The models may use other information on the voter such as his political identification or the issue areas which he cares the most. The information on the articles includes the issue categories, article genre, which candidate the article is about, along with the words in the article body and heading. In addition, the models are provided with information on whether the article is congruent to the voter's personal preference or not.¹⁰

At evaluation time, the models are not given any of the target voter's polling decisions. The model may use other voters' polling data and their profiles, along with their article histories, for the purpose of training. We evaluate the model by leave-one-out cross validation accuracy, precision, and recall on how many decisions were predicted correct.

3.2.4 Proposed works

Motivated reasoning theory emphasizes the effect of continuous update of affective stages on individuals candidate evaluation [86]. This view of voting behavior suggests that the response variables (voter's poll decisions) could be better represented as a sequence. The theory also postulate a set of cognitive stages which control the voter's observable behavior during the campaign. Types of stochastic model which naturally express these assumptions are so called dynamic Bayesian networks. The term refer to frameworks which explicitly encode temporal dependencies in the model. Our proposed work will start with the most straightforward model in this family, hidden Markov model.

⁹The manipulation is automatically generated by the system based on each subjects political outlooks collected during the preliminary interviews. The rate of manipulated articles were varied among the users, to allow controlled data explorations in the later stage.

¹⁰This means that whether the content of the article agrees with what he believes or not. The simulation mechanism on the backend analyzed the individual's profile and automatically generated such manipulative articles during the user studies. See the detail of the process in the original articles [86]

Description
Article topics (finer categorization)
Article topics (coarser categorization)
Distance from the polling n ; $n \in [1..6]$
Article is manipulated or not
Article is intended to be incongruent or not
Article is about the liked candidate, disliked candidate, or other

Table 3.8: Article attributes. Each article is labeled by the conjunction of those attributes.

Given a sequence of articles and a sequence of poll results to predict, we could use various learning framework to represent this problem. One way is to treat the poll results as a response variables, and design a distribution over them conditioning on the trace of articles. Another way is to treat the polling decision as the *cause*, which guides the voter to pick certain articles, then design a joint generative model over both variables. Prediction of the poll in this case is performed as a probabilistic inference on this joint model. We experimented with both directions in the previous tasks.

The starting point of this third task is the second type of approach, but with a little twist. Recall that previously we described the very basic of our comment prediction models as the 0th order Markov model. In this task we represent the sequence of articles and poll results as an instance of the first order hidden Markov model. We propose two sets of experiments:

- Variations of hidden Markov models
- Feature-based hidden Markov models

We will discuss each of them in turn in the following sections.

Representing states and articles

We have not yet specified how articles are represented in this model. In natural language research, the component distributions of HMM are usually categorical distributions. For the moment we will use Bernoulli distribution over “sustained” or “flipped” for $p(y_i|y_{i-1})$ (transition probability), and a categorical distributions over article for $p(x_i|y_i)$ (emission probability). We define here the categories be a set of conjunctives of article attributes. The attributes, conjunctive atoms, include such things as the topic or function of the article (for example issue statement, poll result, or family history), distance from the last poll, whether the article is supposed to be consistent with the subjects political views, and which candidate the articles is about. We “labeled” each article using article metadata and the individual voter’s profiles. See table 3.8 for the detail. We conduct our preliminary experiments with those representation. The largest “vocabulary” of conjunctive labels we used is 1,155.

We run this data representation with support vectortor machine [46] and naive Bayes classifier to test the sanity of this representation, and also to see if we can make reasonable predictions with this dataset. For those experiments, we paired each poll results with all the articles that the subject read up to the poll. Then, we trained each classifier in supervised fashion. We tested the model performance with leave-one-out cross validation. The best performed model is a svm using all the features mentioned in table 3.8, 76.8 % accuracy (f1 = 82.8 %).¹¹ Our baseline method achieved 67.9 % accuracy (f1 = 76.9 %). The baseline is a simple deterministic scheme, which make prediction by keeping the tarry of negative and positive article about the subject’s favorite candidates, and predict “flipped” whenever the negative tarry exceeds the positive ones. The results suggest that this document representation scheme does capture some discriminative signals in our data. We will use this finding as our performance yardstick for the subsequent experiments.

¹¹In this data set, 59.9 % of all votes are “sustained” votes.

Hidden Markov model for voter state prediction

Let's consider each voter independently. Let each campaign article read by this voter be denoted as x_i , and the "state" of the voter at each article as y_i . For now, we will assume the voter state is either "sustained" (i.e., the voter still prefers the originally liked candidate) or "flipped" (the voter no longer prefers the originally liked candidate). Further, we believe that there are stochastic dependencies between both the current state and the article, as well as the current states and the prior state. Thus simplified, an instance of article traces can be represented as a simple trellis: ¹²

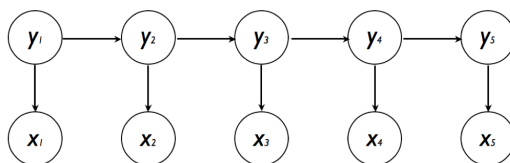


Figure 3.1: A simple hidden Markov model

A type of statistical learning framework which naturally express this form of compound distribution is so called dynamic Bayesian networks [50]. In this task we start with the most straightforward model in this family, hidden Markov model (HMM):

$$p(\mathbf{x}, \mathbf{y}) = p(start) \cdot \prod_i^N p(y_i | y_{i-1}) \cdot p(x_i | y_i) \cdot p(end)$$

Note that the trellis is equivalent to the above joint distribution.

Although HMM is a simple model, it is a convenient, and supported by efficient, well principled inference algorithms. Also, they are quite adaptable to many situation since there are many variations to the basic model.

Experiments with basic HMM

2-state HMM with soft-labeling: HMM can be trained with or without annotation at the state transition. Let's assume that a voter would always vote for his original favorite when he is at the sustained state and against when he is at the flipped state, so we just use the polling information to label each state. Of course, we don't have precise labels at each article. We may use the closest polling result to "softly" label each state (y_i) as either sustained or flipped. Note that in the sanity check experiments in the previous section, we tied articles to the closest voting results in the same manner. This is the simplest form, but with a lot of assumptions. This model can be trained with a maximum likelihood estimate, for which closed form solution is available. The performance measurement from this model will serve as a baseline along with the results from the supervised performance.

Finding tipping points without noisy-labeling: With the soft-labeling above, we need to fix the tipping point in the data arbitrarily. However, we may like to see if we can learn this point from the data, or in another words, see if there is enough signal in the data to converge on the tipping points. For this experiments we hold out part or all of polling results from the training data. We may use part of them to fix some of the state labels, while the rest to be freely adjusted during the training. This model does not have closed form solution, but can be trained with iterative solution such as Expectation Maximization algorithm. The prediction of unseen state is essentially the decoding. For which we can use standard algorithms (e.g., Viterbi decoding).

Finding tipping states: In above two experiments, we fixed the number of states to be two, sustained or flipped. This is in a sense faithful to what the theory suggests. However, the tipping "point" might be longer than a single

¹²We will update this simple assumption later in the extension work section

point. The voter may enter a type of state vacillating between the conflicting emotions toward the candidate. Such a pass may better be captured by using 1 or more state between the two states. Adding a state to the HMM does not make training or decoding more complicated than above 2 states, though makes it more expensive. Also, we will need some mapping scheme between the in-between states and one of the voting actions at the decoding (prediction) time. For the first round, we may simply map in-between states to the historically more copious states. More intelligent mapping schemes are also possible. Note that the same problem arises in other tasks (such as POS tagging) when unsupervised HMM is specified with more states than the presumed target labels.

Constraining transition probability, Higher order Markov model: By construction of our user study, most of the user’s state progression can be assumed to be toward the flipping state. Unconstrained estimation on transition probability may assign some chance of going back from the flipped state to the sustained state. We may like to manipulate the training to prevent such situation. Adding influence from the state further back beyond the previous state is another idea to enrich the state transition process.

Bayesian HMM: Constraints on transition probability were easy to factor into the training algorithm. However, it may make the model too brittle. An alternative way to guide the model learning is to use a prior over the transition probabilities. The Dirichlet distribution is a natural choice given the categorical distributions. Then, constraint (or guidance) on the transition probability can be encoded in the hyper parameters for the Dirichlet distribution, resulting in more flexible guidance than hard constraint. The technique was used for part-of-speech prediction tasks in [37] with significant improvement. Similar techniques have been used successfully in other domains, such as in [43, 28].

Beyond basic HMM

Feature based HMM: In the previous section we specified the article representation as a categorical distribution, and the emission probabilities to be multinomial distributions conditioned on the state. This is a perfectly reasonable choice. However, this particular choice of text representation, somehow limits what we can find from the learned model. For example, it is possible that there is some particularly salient *trait* shared among many articles. Such trends are difficult to see conclusively since the current text representation maps one article to just one type. Even if there are two articles sharing the same traits (for example articles about biographical info and anecdotal accounts are both about personality of the candidate) the commonality between the two articles are lost in our categorical representation. What we need is a flexible distribution which allows each emission to be associated to more than one trait. The issue can be effectively addressed by a feature based text representation and log-linear modeling, similar to the one we employed in the second task. In this experiment we will replace the multinomial emission probability distribution with a locally normalized log-linear distribution of the form (we let w , be the model parameters, $g()$ be the feature functions as in chapter 2):

$$p(x|y) = \frac{\exp(w \cdot g(x, y))}{\sum_{x'} \exp(w \cdot g(x', y))}$$

The change does not require any alternation on the trellis structure above. Note that the denominator (partition function) is a summation over x' , making efficiency issues a potential problem. This type of feature based HMM are recently studied in [9]. The study applied the model to part of speech tagging with HMM successfully. They showed that Expectation Maximization algorithm from the basic HMM parameter learning is in principle still applicable to this case, with gradient based M step. They also introduced a more efficient direct gradient based optimization methods. The similar approach to the optimization was also applied recently in [22].

Possible extensions

Other response values to predict: The implicit hypothesis we are working under in our study is that the changes in the voter’s mental states are driven by the affects which he perceives during the campaign. And those mental changes cause the voter to behave in a certain way (such as flipping from his originally liked candidate). We model textual features (the content of the articles) and their relation to the outcome (e.g., preference change) in our prediction models because we believe that texts are the medium which actuate the emotion in the voters.

Candidate preference however, is not all the behavioral choices influenced by the affects. For example, the shift in affective stage may change the way the voter conducts his information *seeking* (browsing) activities. Predicting what the *next* articles that the voter clicks on could be another interesting challenge. Studies notes that the voters who has been in a motivated reasoning stage tend to go out of their way to find evidences which *refute* (mounting) evidence against their bias [53, 86]. Note that the generative sequence modeling (e.g., HMM), we presented in this task can presumably be able to give such prediction via a standard probabilistic inference.

Another possible modeling task is the prediction on how voters *share* the articles among each other. There is much debate among social science researchers on how affects influence the people's information sharing and forwarding behavior [88]. Do people share more when the information is trusted or infuriating? This is an important question even outside of the theoretical interest, since it directly relates to how affect-loaded information propagates in networks, which in tern influence how the network structure is shaped. In our new DPTE user experiment, which takes place concurrent to the work proposed here, we gather the subject's article forwarding activities among other things.¹³ We may explore those possible other prediction tasks as the extension.

Conditional Random Field (CRF) with latent state variables : A feature-based emission distribution adds more flexibilities to the model. It also makes the model potentially more useful, since the model can now give an explanation on what individual characteristics in the articles affected the process. As we advocated in previous task, we can encode many insights through feature function design. However, there is some limitation on this flexibility; feature functions are limited to only the local variables. In the case of our HMM, only at one article at time. Features involved in the factors from neighboring articles or states can not be encoded in this formalism. For example, having the article, or some traits in the articles, to influence directly to the next states is potentially interesting, though is not possible. In supervised setting, Conditional Random Field is often used to address this type of dilemma in sequence labeling or other structured output problems [95]. One problem of CRF is its costly training. The problem is even more severe for unsupervised learning. Unsupervised learning of CRF has been studied for many tasks, though in many cases their solution exploits particularities in the data or learning situation [92, 40].¹⁴ At the time of this proposal we leave investigation into this possibilities for the extension work.

3.2.5 Conclusion and future work

In this chapter we presented the third task in our dissertation work, the building of statistical model which predicts a person's voting preference change from a trace of his media browsing history. We motivated our task, defined our approach, and outlined the proposed works which we will deliver in the final publication.

Upon completion of this task we would achieve the following contributions:

1. Design of a statistical prediction models built upon social theories.
2. Empirical support to the development of political science theories.
3. Application of dynamic Bayesian networks to text-driven response prediction from the temporally dependent data.

In this proposal we have reported some of the completed works as well as the development sketch of our planned works. As of this proposal, the first phase of data collection, preliminary data analysis, basic model designs were completed. Implementation of the model, model evaluation, model refinement, and model extension are currently in progress. We expect more user data from our collaborators in this coming year. See the next chapter for the milestones and delivery schedule.

¹³The results from user experiments will be published in this coming American Political Science Association's annual meetings. <http://www.apsanet.org/>

¹⁴the locally normalized model such as above feature HMM is often used to avoid this costliness in CRF.

In our immediate course of action, we will first experiment with the basic HMM model and its variations we presented in the section 3.2.4. We then will proceed to more sophisticated models discussed in the extended model section 3.2.4. We will also experiments with other types of response prediction tasks within our DPTE project, applying the techniques we have experimented in this work as well as in the other tasks. In both of those phases we will evaluate our models by performance comparison among the different models. We will also examined the learned models and contrast the findings with the perspectives provided by the political decision theories. We plan to publish our experimental results in relevant computer science conferences as well as at political science or political behavioral conference. Our dissertation will include comprehensive reports.

Chapter 4

Timeline

Project	Work description	Start data –Target data
Congressional bill survival prediction	Model implementation and evaluation	09 – 11/2011
	Model refinement	10 – 12/2011
	Extension	11/2011 – 01/2012
	Publish paper	11/2011 – 01/2012
Polling decision prediction	Model implementation and evaluation	02 – 04/2012
	Model refinement	03 – 05 /2012
Summer internship (05 – 08/2012)		
Polling decision prediction	Extension	09 – 12/2012
	Publish paper	10 – 12/2012
Thesis	Writing	12/2011 – 02/2013
	Diffence	02/2013 (Target)

Table 4.1: Delivery Schedule

Bibliography

- [1] L. Adamic and N. Glance. The political blogosphere and the 2004 U.S. election: Divided they blog. In *Proc. of the 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2005.
- [2] E. Scott Adler and John Wilkerson. The scope and urgency of legislation: Reconsidering bill success in the house of representatives. In *Proc. of annual meetings of the American Political Science Association*, Washington, DC, 2005.
- [3] Nitin Agarwal, Huan Liu, Lei Tang, and Philip S. Yu. Identifying the influential bloggers in a community. In *Proc. of WSDM*, 2008.
- [4] Amr Ahmed and Eric P. Xing. Staying informed: supervised and semi-supervised multi-view topical analysis of ideological perspective. In *Proc. of EMNLP*, 2010.
- [5] William D. Anderson, Janet M. Box-Steffensmeier, and Valeria Sinclair Chapman. The keys to legislative success in the u.s. house of representatives. *Legislative Studies Quarterly*, pages 357–386, 2003.
- [6] Ramnath Balasubramanyan, William W. Cohen, Doug Pierce, and David P. Redlawsk. What pushes their buttons? predicting comment polarity from the content of political blog posts. In *Proc. of Workshop on Language in Social Media (LSM 2011)*, 2011.
- [7] Mikhail Bautin, Lohit Vijayarenu, and Steven Skiena. International sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2008.
- [8] Edward Benson, Aria Haghighi, and Regina Barzilay. Event discovery in social media feeds. In *Proc. of ACL*, 2011.
- [9] Taylor Berg-Kirkpatrick, Alexandre Bouchard-Cote, John DeNero, and Dan Klein. Painless unsupervised learning with features. In *Proc. of NAACL*, 2010.
- [10] D. Blei and M. Jordan. Modeling annotated data. In *Proc. of SIGIR*, 2003.
- [11] D. Blei and J. Lafferty. Topic models. In A. Srivastava and M. Sahami, editors, *Text Mining: Theory and Applications*. Taylor and Francis, In press.
- [12] D. Blei and J. McAuliffe. Supervised topic models. In *Advances in Neural Information Processing Systems 20*, 2008.
- [13] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- [14] S. R. K. Branavan, H. Chen, J. Eisenstein, and R. Barzilay. Learning document-level semantic properties from free-text annotations. In *Proc. of ACL-08: HLT*, 2008.
- [15] Paul Burstein, Shawn Bauldry, and Paul Froese. Bill sponsorship and congressional support for policy proposals, from introduction to enactment or disappearance. *Political Research Quarterly*, pages 295–302, 2005.

- [16] Parliamentarian Charles W. Johnson. How our laws are made, 2003. U.S. House of Representatives Document 10893.
- [17] Paula Chesley, Bruce Vincent, Li Xu, and Rohini Srihari. Using verbs and adjectives to automatically classify blog sentiment. In *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pages 27–29, 2006.
- [18] Joshua Clinton, Simon Jackman, and Doug Rivers. The statistical analysis of roll-call data. *American Political Science Review*, pages 355–370, 2004.
- [19] Lauren Cohen and Christopher Malloy. Friends in high places, 2010.
- [20] D. Cohn and T. Hofmann. The missing link—a probabilistic model of document content and hypertext connectivity. In *Neural Information Processing Systems 13*, 2001.
- [21] Gary W. Cox and Keith T. Poole. On measuring partisanship in roll-call voting: The u.s. house of representatives, 1877-1999. *American Journal of Political Science*, pages 477–489, 2002.
- [22] Dipanjan Das and Slav Petrov. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proc. of ACL*, 2011.
- [23] Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. Probabilistic frame-semantic parsing. In *Proc. of NAACL*, 2010.
- [24] Laura Dietz, Steffen Bickel, and Tobias Scheffer. Unsupervised prediction of citation influences. In *In Proceedings of the 24th International Conference on Machine Learning*, pages 233–240, 2007.
- [25] Peter Sheridan Dodds and Christopher M. Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *JOURNAL OF HAPPINESS STUDIES*, 11(4):441–456, 2008.
- [26] M. Dredze, H. M. Wallach, D. Puller, and F. Pereira. Generating summary keywords for emails using topics. In *Proc. of the 13th International Conference on Intelligent User Interfaces*, 2008.
- [27] Chris Dyer, Jonathan H. Clark, Alon Lavie, and Noah A. Smith. Unsupervised word alignment with arbitrary features. In *Proc. of ACL*, 2011.
- [28] Jaocb Eisenstein and Regine Barzilay. Bayesian unsupervised topic segmentation. In *Proc. of EMNLP*, 2008.
- [29] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed membership models of scientific publications. *Proc. of the National Academy of Sciences*, pages 5220–5227, April 2004.
- [30] Lawrence C. Evans. Participation and policy making in senate committees. *Political Science Quarterly*, 1991.
- [31] Katja Filippova and Keith B. Hall. Improved video categorization from text metadata and user comments. In *Proc. of SIGIR*, 2011.
- [32] Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2 2010.
- [33] Sean Gerrish and David Blei. A language-based approach to measuring scholarly impact. In *Proc. of ICML*, 2010.
- [34] Sean Gerrish and David Blei. Predicting legislative roll calls from text. In *Proc. of ICML*, 2011.
- [35] Eric Gilbert, Tony Bergstrom, and Karrie Karahalios. Blogs are echo chambers: Blogs are echo chambers. In *Proc. of HICSS*, 2009.
- [36] Namrata Godbole, Manjunath Srinivasaiah, and Steven Skiena. Large-scale sentiment analysis for news and blogs. In *Proc. of ICWSM*, 2007.

- [37] Sharon Goldwater and Thomas L. Griffiths. A fully bayesian approach to unsupervised part-of-speech tagging. In *Proc. of ACL*, 2007.
- [38] Justin Grimmer. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*, Forthcoming.
- [39] D. Gruhl, R. Guha, David Liben-nowell, and A. Tomkins. Information diffusion through blogspace. In *In WWW 04*, pages 491–501. ACM Press, 2004.
- [40] Aria Haghighi and Dan Klein. Prototype-driven learning for sequence models. In *Proc. of ACL*, 2006.
- [41] Richard Hall. Participation and purpose in committee decision-making. *American Political Science Review*, pages 105–128, 1998.
- [42] Thad Hall. *Authorizing Policy*. Ohio State University Press, 2004.
- [43] Zhang Hao, Chris Quirk, Robert C. Moore, and Daniel Gildea. Bayesian learning of non-compositional phrases with synchronous parsing. In *Proc. of ACL-08*, 2008.
- [44] Gregor Heinrich. Parameter estimation for text analysis. Technical report, University of Leipzig, 2008.
- [45] Simon Jackman. Multidimensional analysis of roll call data via bayesian simulation. In *Proc. of NAACL*, pages 227–241, 1991.
- [46] Thorsten Joachims. Making large-scale support vector machine learning practical. *Advances in kernel methods: support vector learning*, pages 169–184, 1999.
- [47] Mahesh Joshi, Dipanjan Das, Kevin Gimpel, and Noah A. Smith. Movie reviews and revenues: An experiment in text regression. In *Proc. of NAACL*, 2010.
- [48] David Karpf. Understanding blogspace. *Journal of Information Technology and Politics*, 5(4):369–385, 2008.
- [49] A. Kittur, B. Suh, and E. Chi. What’s in wikipedia? mapping topics and conflict using collaboratively annotated category links. In *Proc. of CHI*, 2009.
- [50] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.
- [51] Glenn Krutz. Issues and institutions: Winnowing in the u.s. congress. *American Journal of Political Science*, 49:313–26, 2005.
- [52] Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pages 100–107, 2006.
- [53] Ziva Kunda. The case for motivated reasoning. *Psychological Bulletin*, 108:480 – 498, 1990.
- [54] Michael Laver, Kenneth Benoit, and John Garry. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97:311–331, 2003.
- [55] Michael Laver and John Garry. Estimating policy position from political texts. *American Journal of Political Science*, 44(3), 2000.
- [56] Eric Lawrence, John Sidesa, and Henry Farrell. Self-segregation or deliberation? blog readership, participation, and polarization in american politics. *Perspectives on Politics*, 4(8):141–157, 2010.
- [57] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *Proc. of KDD*, 2007.

- [58] Jure Leskovec, Mary Mcglohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. Cascading behavior in large blog graphs. In *In SDM*, 2007.
- [59] W.-H. Lin, E. Xing, and A. Hauptmann. A joint topic and perspective model for ideological discourse. In *Proc. of 2008 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2008.
- [60] R. Malouf and T. Mullen. Graph-based user classification for informal online political discourse. In *Proc. of the 1st Workshop on Information Credibility on the Web*, 2007.
- [61] Donald R. Matthews. The foldways of the united states senate: Conformity to group norms and legislative effectiveness. *American Political Science Review*, 59:1064–1089, 1959.
- [62] A. McCallum. Multi-label text classification with a mixture model trained by EM. In *AAAI Workshop on Text Learning*, 1999.
- [63] Scott McClurg. Social networks and political participation: The role of social interaction in explaining political participation. *Political Research Quarterly*, 64(3):480 – 498, 2003.
- [64] Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. Structured models for ne-to-coarse sentiment analysis. In *Proc. of ACL*, 2007.
- [65] D. Mimno and A. McCallum. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Proc. of UAI*, 2008.
- [66] Gilad Mishne and Natalie Glance. Leave a reply: An analysis of weblog comments. In *Proc. of Workshop on the Weblogging Ecosystem*, 2006.
- [67] Burt Monroe, Michael Colaresi, and Kevin M. Quinn. Fightin’ words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, pages 372–403, 2008.
- [68] T. Mullen and R. Malouf. A preliminary investigation into sentiment analysis of informal political discourse. In *Proc. of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.
- [69] R. Nallapati and W. Cohen. Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs. In *Proc. of the 2nd International Conference on Weblogs and Social Media*, 2008.
- [70] Brendan O’Connor, Jacob Eisenstein, Eric P. Xing, and Noah A. Smith. Discovering demographic language variation. In *Proc. of NIPS Workshop on Machine Learning for Social Computing*, 2010.
- [71] Walter Oleszek. Congressional procedures and the policy process. *Congressional Quarterly*, pages 105–128, 2001.
- [72] Iadh Ounis, Maarten de Rijke, Craig Macdonald, Gilad Mishne, and Ian Soboroff. Overview of the TREC-2006 Blog Track. In *Proceedings of the 15th Text REtrieval Conference (TREC 2006)*, 2006.
- [73] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proc. of ACL*, 2004.
- [74] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2:1–135, January 2008.
- [75] Souneil Park, Minsam Ko, Jungwoo Kim, Ying Liu, and Junehwa Song. The politics of comments: predicting political orientation of news stories with commenters’ sentiment patterns. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, 2011.
- [76] Slav Petrov and Dan Klein. Discriminative log-linear grammars with latent variables. In *Proc. of NIPS 20*, 2008.

- [77] Keith Poole and Howard Rosenthal. Spatial model for legislative roll call analysis. *American Journal of Political Science*, pages 357–384, 1985.
- [78] Keith Poole and Howard Rosenthal. Patterns of congressional voting. *American Journal of Political Science*, pages 118–178, 1991.
- [79] Hoifung Poon and Pedro Domingos. Unsupervised semantic parsing. In *Proc. of EMNLP*, EMNLP '09, 2009.
- [80] David Price. *Who Makes the Laws? Creativity and Power in Senate Committees*. Transaction Publishers, 1972.
- [81] Kevin M. Quinn. Bayesian factor analysis for mixed ordinal and continuous responses. *Political Analysis*, pages 338–353, 2004.
- [82] Kevin M. Quinn, Burt Monroe, Michael Colaresi, Michael Crespin, and Drago Radev. How to analyze political attention with minimal assumptions and costs. *American Journal of Political*, pages 209–228, 2010.
- [83] Kevin M. Quinn, Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. An automated method of topic-coding legislative speech over time with application to the 105th–108th U.S. Senate, 2006. Midwest Political Science Association Meeting.
- [84] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proc. of EMNLP*, 2009.
- [85] David P. Redlawsk. Hot cognition or cool consideration? testing the effects of motivated reasoning on political decision making. *The Journal of Politics*, pages 1021–1044, 2002.
- [86] David P. Redlawsk, Andrew J. W. Civettini, and karen M. Emmerson. The affective tipping point: Do motivated reasoners ever “get it”? *Political Psychology*, 31(4):563–593, 2009.
- [87] David P. Redlawsk and Richard R. Lau. Understanding individual decision making using process tracing. In *Proc. of European Consortium for Political Research*, 2009.
- [88] David P. Redlawsk, Doug Pierce, William W. Cohen, Tae Yano, and Ramnath Balasubramanyan. Assessing the effects of emotion-laden messages in a social network. In *Proc. of American Political Science Association*, 2011.
- [89] Alan Ritter, Colin Cherry, and Bill Dolan. Unsupervised modeling of twitter conversations. In *Proc. of NAACL*, 2010.
- [90] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P Smyth. The author-topic model for authors and documents. In *Proc. of the 20th Conference on Uncertainty in Artificial Intelligence*, 2004.
- [91] Noah A. Smith. *Linguistic Structure Prediction*. Morgan and Claypool, 1st edition, May 2011.
- [92] Noah A. Smith and Jason Eisner. Contrastive estimation: Training log-linear models on unlabeled data. In *Proc. of ACL*, 2005.
- [93] M. Steyvers and T. Griffiths. Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, editors, *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum, 2007.
- [94] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. L. Griffiths. Probabilistic author-topic models for information discovery. In *Proc. of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004.
- [95] Charles Sutton and Andrew McCallum. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 2011. To appear.
- [96] Sakaki Takeshi, Okazaki Makoto, and Matsuo Yutaka. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proc. of WWW*, 2010.

- [97] I. Titov and R. McDonald. A joint model of text and aspect ratings for sentiment summarization. In *Proc. of ACL-08: HLT*, 2008.
- [98] Kevin Wallsten. Political blogs: Transmission belts, soapboxes, mobilizers, or conversation starters? *Journal of Information Technology and Politics*, 4(3):19–40, 2008.
- [99] Tae Yano, William W. Cohen, and Noah A. Smith. Predicting response to political blog posts with topic models. In *Proc. of NAACL-HLT*, 2009.
- [100] Tae Yano and Noah A. Smith. What’s worthy of comment? content and comment volume in political blogs with topic models. In *Proc. of ICWSM*, 2010.
- [101] Sarita Yardi and Danah Boyd. Tweeting from the town square: Measuring geographic local networks. In *Proc. of ICWSM*, 2010.
- [102] Ainur Yessenalina, Yisong Yue, and Claire Cardie. Multi-level structured models for document-level sentiment classification. In *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2010.
- [103] Dani Yogatama, Michael Heilman, Brendan O’Connor, Chris Dyer, Bryan R. Routledge, and Noah A. Smith. Predicting a scientific community’s response to an article. In *Proc. of EMNLP*, 2011.
- [104] Omar F. Zaidan and Jason Eisner. Modeling annotators: a generative approach to learning from annotator rationales. In *Proc. of EMNLP*, 2008.
- [105] Omar F. Zaidan, Jason Eisner, and Christine Piatko. Using ”annotator rationales to improve machine learning for text categorization. In *Proc. of NAACL*, 2007.
- [106] H. Zhang, B. Qiu, C. L. Giles, H. C. Foley, and J. Yen. An LDA-based community structure discovery approach for large-scale social networks. In *Proc. of the IEEE International Conference on Intelligence and Security Informatics*, 2007.
- [107] Yan Zhang, A.J. Friend, Amanda L. Traud, Mason A. Portes, James H. Fowler, and Peter J. Mucha. Community structure in congressional cosponsorship networks. *Physica A*, 387(7), 2007.
- [108] Jun Zhu, Ahmed Amr, and Eric P. Xing. MedLDA: maximum margin supervised topic models for regression and classification. In *Proc. of ICML*, 2009.