Discovering Voter Preferences in Blogs using Mixtures of Topic Models

Pradipto Das
University at Buffalo
201 Bell Hall
Buffalo, NY 14228
pdas3@buffalo.edu

Rohini Srihari
University at Buffalo
201 Bell Hall
Buffalo, NY 14228
rohini@cedar.buffalo.edu

Smruthi Mukund
University at Buffalo
201 Bell Hall
Buffalo, NY 14228
smukund@buffalo.edu

ABSTRACT

In this paper we propose a new approach to capture the inclination towards a certain election candidate from the contents of blogs and to explain why that inclination may be so. The method is based on the availability of "ground truth" speeches from the election candidates that are labeled and also on the collection of noisy blogs which are not labeled in any way. In this unsupervised learning scenario, we used probabilistic topic models to cluster the ground truth documents for each candidate into different underlying latent themes. The same topic models were then applied on the blog collection and the "orientation" of each of the blogs with different themes of the election candidate speeches was performed using KL divergence of the topic distribution over the overlapping vocabularies. We used four models for such theme matching, one with a baseline topic model and the other three by weighting the baseline topic model with the positive, negative and the neutral sentiments of the topics. We then used a collaborative objective function to combine the outcome of candidate preference for the blogs under the four models using an Expectation Maximization algorithm. The novelty of our method is highlighted in its use of unannotated data as well as in the combination of the views of the different "experts" explaining the same phenomenon.

Categories and Subject Descriptors

I.5.1 [**Pattern Recognition**]: Models—*statistical*

General Terms

Topic Models, Social network, blogs, KL Divergence

1. INTRODUCTION

In this paper, we attempt to investigate the problem of how blogger inclinations towards election candidates can be identified from blogs. By blogger inclinations, we highlight the fact that if two candidates are running for presidency in a presidential election campaign, then if people were blogging on their campaign promise speeches, were they really doing so because they like the themes spoken by the candidates.

Topic models [4] have become the cornerstone for understanding the thematic organization of large text corpora in an unsupervised fashion. These models define probabilistic generative process for document generation in a robust manner. For the sake of brevity we do not iterate over the technicalities of such topic models. Interested readers are referred to Latent Dirichlet Allocation (LDA) and Correlated Topic Models (CTM) [4, 2]. We note that in the original version of the topic model, LDA, the proportions of the topic mixtures in a document are weakly correlated due to the multinomial nature of the topic proportions in documents. Further, modeling correlation between topics using logistic normal priors was posed and solved in [2]. It turned out that this model, conveniently referred to as the Correlated Topic Model (CTM), had much lower document perplexity. To this end, we adopt the topic modeling framework owing firstly to identify the latent themes in the speeches of the election candidates. These themes are nothing but the topics and hence the distributions over the vocabulary of respective candidates. In our case the candidates are Senator Obama and Senator McCain and we manually collected their transcribed speeches and news reports through simple Google searches. These sets of speeches are referred to as ground truths through out this paper and goal was to discover why bloggers may be oriented towards one candidate vs. the other. The ground truth documents can be found at the author's webpage[1].

A similarity between the new blog post and the modeled documents can be measured by computing the topic model document likelihoods. But using likelihoods as a comparitive measure is not suitable as blogs really don't represent "held-out" test set. We thus used the same topic model with the same number of topics over the ground truth data and the blog post collection. We then compared the average KL divergence of the overlapping word histograms (probability densities) over topics for each blog post against those in the two sets of ground truths. The *orientation* of the blog post as being pro-Obama or pro-McCain was made based on the low divergence.

The problem of identifying blogger political inclinations become incredibly hard due to the presence of sarcasms. Sarcastic blogs tend to have similar "topic" distributions over

[1] www.buffalo.edu/~pdas3

AND '09, July 23-24, 2009, Barcelona, Spain

words with completely different tone towards the person or object being discussed. To combat the effects of sarcasms statistically, we defined four models - one was the baseline CTM topic model; the other three were three different positive, negative and objective models which just weighed the probability of each word under a topic with the posterior probability of the respective positive, negative and objective sentiments for that topic. We finally define a collaborative objective function for the models to select the label of a blog document of only that model which had the maximum confidence in it's labeling. Note that the use of topic models also helps us to explain the reason for the inclinations in blogs towards different candidates. Note that we are not interested in the positive or negative opinion of the blogger towards an election candidate, but explaining why it may be so based on theme matching.

The paper is organized as follows. The next two section reviews some related work and a brief introduction to probabilistic topic model, followed by the mention of how the Spinner.com data was adapted for this task and highlight the indexing and data cleaning. Then we review some basic concepts on topic models and discuss the techniques of our proposed method. We then follow up with results and analysis of the output of the proposed method. The paper is concluded in the last section with some ideas for improvement in future work.

2. RELATED WORK

The wide coverage of the recent presidential campaign through the use of electronic media has seen a surging interest in analysis of data from social websites. A very recent commercial opinion utility product called *Jodange*[2] attempts to "automatically extract opinions from online content in order to understand sentiment." This is related to extraction of top opinions w.r.t. a specified topic (not the same "topic" from an unsupervised topic model) or an opinion holder and a sentiment score for those opinions. Yejin et al.[6] present a novel learning-based approach that incorporates structural inference motivated by compositional semantics into the learning procedure. The goal of learning is typically to learn the sentiment polarity class posterior given complete(labeled) data in the form of blog posts with sentiments and use a classifier like Support Vector Machine to classify sentiment polarity of new blogs. These methods use blog content. However, annotated data is required to train the classifier. A similar approach is taken by [5] where Conditional Random Field approach is used for classification. With annotated data, supervised classification remains the method of choice which is further supported in the works of [8] and [14] that encompass even methods like Naive Bayes classification.

There has been some work in modifying a fundamentally unsupervised data label learning algorithm like topic model to fit supervised learning scenarios. The most prominent of them being the Supervised Topic Model (sLDA) [3]. However the goal of sLDA was to predict labels like movie ratings based on the text of the reviews. Good predictive topics should differentiate ratable words like "thumbs-up", "thumbs-down", and "neutral," without regard to genre. But topics estimated from an unsupervised model may correspond to genres, if that is the dominant structure in the corpus. Another work authored by [12] and explores a semi-supervised learning algorithm for classifying political blogs in a blog network and ranking them within classes. The blogs were classified as democratic and republican blogs and their method used only the link structure of blog sites and an adaptation of the well known PageRank algorithm to endorse links as Republican or Democratic. Similar link analysis to determine the hub and authority scores of the blogs had been performed by [1]. We cannot apply link models on the ground truth campaign speech data due to the lack of any link structure in the documents whatsoever.

[2]www.jodange.com

3. BACKGROUND FOR OUR METHOD

In this section we give a brief introduction to topic models and point out a few relevant facets of such modeling that was useful to implement our method.

3.1 Topic Models

Topic models are robust statistical models that tries to explain textual document generation process. One such model, CTM is shown as a directed graphical model in Figure 1. Graphical models allow us to graphically represent interactions between unobserved, observed and parameters in a clear and concise way using directed arcs for causality relationships and sqaure plate notation for identifying variable repetition.

In Figure 1 we observe that the only observed random variable is the word label with N being the number of word positions in document d. $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}$ are the model parameters and $\boldsymbol{\eta}$ represents the topic or theme proportions for a document. Clearly, the use of proportions suggest that the only input to the model are documents represented in terms of their word counts. The likelihood of a document (which is a vector of the same size as the corpus vocabulary) w.r.t model parameters is given as

$$\int d\boldsymbol{\eta}_d \, p(\boldsymbol{\eta}_d|\mu,\Sigma)(\prod_{n=1}^{N_d}\sum_{z_{d_n}} p(z_{d_n}|\boldsymbol{\eta}_d)p(w_{d_n}|z_{d_n},\boldsymbol{\beta}))$$

This likelihood is intractable to compute and approximate algorithms like Variational Bayes are used to overcome the intractibility. To explain the phenomenon catured by the

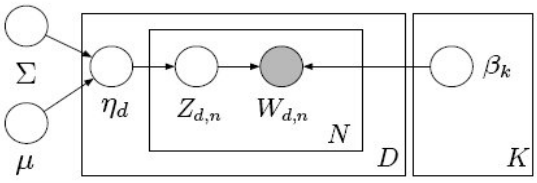

Figure 1: The graphical model for the Correlated Topic Model

graphical model representation in Figure 1, we write the document generation process in the following form. Let $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ be a K-dimensional mean and covariance matrix, and let topics $\boldsymbol{\beta}_{1:K}$ be K multinomials over a fixed vocabulary of words. This K is the same as the number of topics. The

correlated topic model assumes that an N-word document d arises from the following generative process:

1. Draw $\boldsymbol{\eta} \mid \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
2. For $n \in \{1, ..., N\}$:
   (a) Draw topic assignment $z_n \mid \boldsymbol{\eta}$ from $Mult(f(\boldsymbol{\eta}))$
   (b) Draw word $w_n \mid \{z_n, \boldsymbol{\beta}_{1:K}\}$ from $Mult(\boldsymbol{\beta}_{z_n})$

with $f(\eta_i) = \frac{exp(\eta_i)}{\sum_j exp(\eta_j)}$ with $i \in \{1, ..., K\}$ and $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a normal distribution with mean $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

The basic idea behind topic models is quite simple. Assume that you have K faced dice with the K numbers on the face denoting the K latent themes or topics (distributions over the vocabulary) for the entire corpus. The corpus is essentially a vocabulary V of words and each document is a sequence of N positions with each position filled by choosing a word from the vocabulary. Thus the vocabulary is represented by a V faced dice and there are K of those - one for each topic.

A document is thus generated in the following way: For each word position in the document, roll the K faced dice to choose a topic number and for that topic number roll the corresponding V faced dice and fill up the word position with the word from the vocabulary indexed by the number on the face of the dice.

The important thing to note here that there are probabilities associated with each number on the face of all die. These probabilities are based on assumed statistical distributions whose parameters are learned through different algorithms.

It is quite important to note that the textitnumbers on the faces of the K-sided dice doesn't mean much by themselves. For example let us consider two separate document collections with one as a collection of essays on "war and peace"($collection_1$) and another as a collection of news reports on "the war in Iraq"($collection_2$). If we assume the K-faced dice to be a coin i.e. K=2, then ideally the two topics for the former collection would contain words related to war and peace respectively and that for the latter collection will consist of words related to war and not war like President Bush's foreign policy, intelligence reports, biological weapons etc. respectively. Imagine that for $collection_1$, topic K_1 is on war and topic K_2 is on peace and that for $collection_2$, topic K_1 is on something other than war and topic K_2 is on war. Clearly, the topic on war for both collections would contain similar word distributions about the topic "war". We thus observe that even if the same topic model is used to model two different collections, the correspondence between the topics of different document collections is not easily captured. Furthermore, the assignment of these topic numbers is purely arbitrary and this can be easily verified manually by plotting the distributions of the topics over the vocabulary for the different document collections. This fact became very useful when we compared the KL divergences of the blog post w.r.t the ground truth document collections.

The intial K faced dice represents topic proportions in documents with the topics themselves being multinomial distributions over the vocabulary. The original LDA topic model uses a Dirichlet distribution with only a single parameter to model topic proportions, $\boldsymbol{\eta}$ in Figure 1. However, to capture stronger correlation between these proportions, the topic proportions are drawn from a logistic normal with two parameters rather than a Dirichlet. This correlated topic modeling also showed lower document perplexity than LDA.

4. THE PROPOSED METHOD

In this section we describe our method to discover blogger orientation towards a political campaigner using the Spinner.com[10] blog collection.

4.1 Description of the Dataset

The initial dataset is a collection of 44 million blog posts which span the period between August 1st to October 1st, 2008. The time period spans a number of big news events like the Olympics held in China, the Presidential conventions held by the Democrats and the Republicans as well as the beginning of the financial crisis. Further the entire dataset is arrranged into 14 tiers based on an approximate search engine ranking. To determine the validity of our idea, we have filtered the noisy blog data using Lemur[3] from the tiers 1 through 4. We also collected a few transcribed speeches from both the speakers - Obama and McCain, from the recent Presidential campaign media coverage. Since most speeches revolved around similar themes, we collected only 20 unique documents for McCain and 28 unique documents for Obama after careful review. Some of these were also unbiased news reports of speeches obtained by Googling *"mccain campaign promises"* and *"obama campaign promises"*. The majority of articles for Obama were collected from the website - www.whitehouse.gov.

4.2 Filtering Blog Data

To index the data, the Spinner.com dataset was first converted to the TREC web format. The documents of tier 1 through 4 were indexed using the standard inverted list and tf-idf method and queried to retrieve documents containing keywords {*obama, mccain*} *AND* {*policy*}. The actual blog post is enclosed in the XML tag $< Description >$ and this part of the text was extracted and cleaned to be fed as input to the Correlated Topic Model. Further, a second step of filtering was applied to index the data for the baseline and weighted topic models. In this step a standard list of stopwords and keywords Barrack, Obama, John, McCain, Joe, Biden, Sarah and Palin were removed because these words are not indicative of their policies. Porter stemming was also applied to the list of remaining valid words.

4.3 Applying Topic Models

In the formulation of topic models like LDA[4] and CTM[2] the exact inference of finding a theme or topic given a particular word in a document is intractable. Thus a deterministic variational Bayesian approach is undertaken to find a tight lower bound to the observed words and hence the document likelihood given the model parameters.

[3]www.lemurproject.org

4.4 Finding Themes of Candidate Speeches

Our method exploits the observation that candidates make *promise* speeches belonging to different themes with each theme differing only on some key selection of words. The bloggers show preference towards one candidate vs. the other citing the respective theme with the support of similar or closely related vocabulary. Note that the vocabulary of the bloggers and that of the ground truth speeches may be different but experimentally it is observed that there are quite a few overlaps in key phrases of the themes.

In our experiments, we applied a 42-topic CTM over the two sets of ground truth speeches from both the presidential candidates, Senator Obama and Senator McCain and collected two sets of topic-word parameter matrices $\boldsymbol{\beta}$ with rows summing to 1. The value of 42 was determined by obseving the maximum likelihoods upon inference on each candidate ground truth. Since a document can be looked upon as both a word and topic simplex[4], we converted the global topic distributions over words to word distributions over topics by normalizing the columns of $\boldsymbol{\beta}$ where each entry in column j is just $p(w_j|z_k)$, $k \in \{1, ..., 42\}$ with w_j being the j^{th} word in the vocabulary and z_j being the indicator variable which takes the value 1 when the word w_j is assigned to topic k. We thus get empirical point estimates of $p(z_k|w_j)$ for both the ground truth collections.

Next we define a sentiment weighted topic model to be the base topic model, but where $p(w_j|z_j)$ is replaced by $p(s_{abc}, w_j|z_j)$, in which *abc* can be "**pos**itive", "**neg**ative" or "**obj**ective" sentiments associated with a word. We thus have the following equations for each document in the blog collection:

$$p(s_{pos}, w_{j_{blog}}|z_j, \boldsymbol{\beta}) \propto p(w_{j_{blog}}|z_j, \boldsymbol{\beta})\delta(w_{j_{blog}}, w_{x_{cand}}) \times \times (p(z_j|s_{pos}) \times p_{prior}(s_{pos})) \quad (1)$$

$$p(s_{neg}, w_{j_{blog}}|z_j, \boldsymbol{\beta}) \propto p(w_{j_{blog}}|z_j, \boldsymbol{\beta}) \times \delta(w_{j_{blog}}, w_{x_{cand}}) \times (p(z_j|s_{neg}) \times p_{prior}(s_{neg})) \quad (2)$$

$$p(s_{obj}, w_{j_{blog}}|z_j, \boldsymbol{\beta}) \propto p(w_{j_{blog}}|z_j, \boldsymbol{\beta}) \times \delta(w_{j_{blog}}, w_{x_{cand}}) \times (p(z_j|s_{obj}) \times p_{prior}(s_{obj})) \quad (3)$$

where, $\delta(w_{j_{blog}}, w_{x_{cand}})$ defines the overlap between the j^{th} word in the blog document vocabulary and x^{th} word in a candidate speech vocabulary; $p(w_{j_{blog}}|z_j, \boldsymbol{\beta})$ is the word topic probability under the base topic model; $\delta()$ is the delta function. We note that this set up used for each pair (blog, candidate). While calculating $p(s_{pos}, w_{j_{cand}}|z_j, \boldsymbol{\beta})$ for a particular candidate *cand*, the delta function is always one. We thus have four models - one is the baseline topic model and the rest three are the sentiment weighted baseline topic model. Note that the term $p(w_{j_{blog}}|z_j, \boldsymbol{\beta})$ is assumed to be independent of the sentiment when topic multinomials are calculated. This is indeed a very hard assumption, however, since the z_j's depend on the w_j's and inturn their sentiments, the overall sentiment of the topic is captured partially.

The sentiment for each word is calculated using the widely popular SentiWordNet [9] which contains the sentiment weightage of each word w.r.t positivity, negativity and objectivity. The term s_{pos} represents the positive sentiment and $p_{prior}(s_{pos})$ is the prior distribution of the positive sentiment w.r.t *a* dataset. The respective fractions of the sentiments in the different vocabularies serve as the parameters for the sentiment priors.

The sentiment priors are calculated for each of the ground truth speeches as well as the blog collection. We use the following assumptions:

$$p(z_j = k|s_{pos}) \propto \tau_k^{\sum_{j=1}^{V_{blog}}[w_{k,j} \times \delta(w_{k,j} \in \{w_{k,j}^{pos}\})]} \quad (4)$$

$$p_{prior}(s_{pos}) = Dir(\alpha_{pos}) \quad (5)$$

$$p(s_{pos}|z_j = k) = Dir(\sum_{j=1}^{V_{blog}}[w_{k,j} \times \delta(w_{k,j} \in \{w_{k,j}^{pos}\})] + \alpha_{pos}) \quad (6)$$

$$\alpha_{pos} = \frac{\sum_{j=1}^{V_{blog}} w_j^{(c)} \times \delta(w_j \in \{w_j^{pos}\})}{\sum_{j=1}^{V_{blog}} w_j^{(c)}} \quad (7)$$

$$\tau_k = \frac{\sum_{j=1}^{V_{blog}} w_{k,j} \times \delta(w_{k,j} \in \{w_{k,j}^{pos}\})}{\sum_{j=1}^{V_{blog}} w_{k,j}} \quad (8)$$

where, $p(z_j = k|s_{pos})$ is the probability of topic k for positive sentiment; τ_k is the multinomial parameter for the positive sentiment under the topic k; $\sum_{j=1}^{V_{blog}} w_{k,j}$ is the sum of all words with positive sentiment found in SentiWordNet. This count is zero if the word is not found in SentiWordNet. $w_{k,j}^{pos}$ is the $w_{k,j}^{th}$ word having positive sentiment and $\{w_{k,j}^{pos}\}$ is the set of all such words under the k^{th} topic. We replace "pos" by "neg" and "obj" to compute these quantities for the other models. $p(s_{pos}|z_j)$ is the posterior probability of the positive sentiment given topic $z_j = k$. Dir(.) is the dirichlet distribution and $w_j^{(c)}$ is the count of the j^{th} word.

4.5 Finding Orientation of Blogs towards Candidate Themes

Discovering the orientation of the blog themes to ascertain whether the author(s) of the blog is reflecting the ideas of Obama or McCain was our main objective. Due to lack of quantitative measures for subjectivity, we decided to employ the Kullback-Liebler(KL) divergence measure between two word topic distributions. This also gave us a hard measure about the political candidate orientation of a blogger. Essentially, if $p(w_{blog\_dcoument}|topic)$ is the variational word-topic distribution per blog document and $q^{(cand)}(w_{cand\_speech}|topic)$ is the distribution vector of the columnwise renormalized estimated global topic-word parameter matrix obtained from the ground truth vocabulary of candidate *cand*, then the KL divergence between p and q under the base model CTM, is determined as

$$D_{KL}(p||q^{(cand)}) = \sum_{k=1}^{K}[p(w_{blog\_document}|topic = k) \times \log \frac{p(w_{blog\_dcoument}|topic = k)}{q^{(cand)}(w_{cand\_speech}|topic = k)}] \quad (9)$$

For each word in the blog, we compared its divergence over *inferred* topics for the same word in one candidate's speeches vs. the others. If the divergence score for candidate 1 is less

than that for candidate 2, the vote of the blog towards candidate 1 is incremented by one. The document is labeled by the candidate who received the maximum votes. The final score of a blog post was composed by subtracting the respective number of votes from this divergence. By doing this, longer informative blogs appeared higher in the ranked list. The same measure is employed for the joint distributions of (sentiment,word) pairs over the topics for the respective "positive", "negative" and "objective" sentiment weighted topic models, where $p(w_{blog\_document}|topic = k)$ and $q^{(cand)}(w_{ground\_truth}|topic = k)$ are replaced by

$$p(s_{pos}, w_{j_{blog}}|z_{j_{blog}}, \boldsymbol{\beta}) \qquad and \qquad p(s_{pos}, w_{j_{cand}}|z_{j_{cand}}, \boldsymbol{\beta})$$

respectively. We note that the goal here was to find the orientation of blogs in an *unsupervised* fashion and we only have "labels" on the ground truths speeches.

4.6 Collaborative Model Combination

In this section, we describe how the four models that assign election candidate inclination labels to the blog documents using a mixture of experts [11] and the Expectation Maximization (EM) algorithm [7]. The goal here is firstly to compute a confidence on the labels that each model assigns to the blog documents and then to define an objective function that fosters collaboration between the different models so that the most confident labeling is captured.

In our case, each blog document is represented as vector over V_{Blog}. Thus each of the M documents i is represented as a pair $(\mathbf{x_i}, \mathbf{w_i})$ under each model h, where $\mathbf{x_i}$ is the count vector of term frequencies and $\mathbf{w_i}$ is the KL divergence weights for those terms that are matching with the target candidate ground truth data. The combined weight over all blogs under model h is taken as $w_h^{(v)} = \frac{M}{\sum_{i=1}^{M}(\frac{1}{1+w_1^{(v)}}+\ldots+\frac{1}{1+w_M^{(v)}})}$, where v is an index in the blog vocabulary. The harmonic is computed component-wise and is used since it is much less prone to outliers than the arithmetic mean. The final collaborative objective function is given as:

$$E_i = \frac{\alpha}{2}\sum_{i=1}^{M}(f(\mathbf{x_i}^T\mathbf{w_j}))^2 - \frac{\beta}{2}\sum_{j=1}^{H}\sum_{l=1,l\neq j}^{H} cos^2(\mathbf{w_j}^T\mathbf{w_l}) \quad (10)$$

$$f(\mathbf{x_i}^T\mathbf{w_j}) = (1 + exp(-\mathbf{x_i}^T\mathbf{w_j}))^{-1} \quad (11)$$

where, H is the total number of models. We note here that for each document the weights of each of the terms are the minimum of the KL divergences between the blog and the respective candidate ground truths, they are always ≥ 0 and thus the sigmoid always returns a value ≥ 0.5. The cosine term penalizes models with close to orthogonal weights and supports model weight vectors in close proximity. E_i is the overall cost function of the mixture of experts modeling the i^{th} datum, while $E_{i,j} = \frac{\alpha}{2}(f(\mathbf{x_i}^T\mathbf{w_j}))^2 - \frac{\beta}{2}\sum_{j=1}^{H}\sum_{l=1,l\neq j}^{H} cos^2(\mathbf{w_j}^T\mathbf{w_l})$ is the cost function of the j^{th} expert modeling the i^{th} datum.

Let the hidden indicator variable h_i taking on value j denote the i^{th} datum being modeled by the j^{th} topic model with probability π_j , then the probability of observing datum $\mathbf{x_i}$ given the parameters $\{\mathbf{w_1}, ..., \mathbf{w_H}\} = \boldsymbol{\Theta_w}$ and indicator $h_i = j$ is $P(\mathbf{x_i}|\mathbf{w_j}, h_i = j) \propto e^{-E_{ij}}$

We now use the EM algorithm to automatically choose the expert for the i^{th} datum. The probability of observing M i.i.d. training samples $\{\mathbf{x_i}\}_{i=1}^M$ given the weights $\boldsymbol{\Theta_w}$ and mixing proportions $\boldsymbol{\pi}$ is given by

$$\log P(\{\mathbf{x_i}\}_{i=1}^M|\{\mathbf{w_1}, ..., \mathbf{w_H}\}, \boldsymbol{\pi}) = \prod_{i=1}^{M}[\sum_{j=1}^{H}(\pi_j \times P(\mathbf{x_i}|\mathbf{w_j}, h_i = j))] \quad (12)$$

where each h_i is some hidden variable distributed according to the prior mixing proportions $\boldsymbol{\pi}$. Taking the logarithm of this probability, and introducing a set of variational distributions $Q = \{Q_i(h_i)\}_{i=1}^M$ over the hidden indicators for each of the data points, using Jensen's inequality we obtain a lower bound on the likelihood of the parameters $\boldsymbol{\Theta_w}$ for the model, which we denote by $\mathcal{F}(\boldsymbol{\Theta_w}, \{Q_i(h_i)\}_{i=1}^M)$. The likelihood of the dataset is given by

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Theta_w}) &\equiv \log P(\{\mathbf{x_i}\}_{i=1}^M|\boldsymbol{\Theta_w}) \\ &\geq \sum_{i=1}^{M}\sum_{j=1}^{H} Q_i(h_i)\log(\frac{P(h_i)P(\mathbf{x_i}|\mathbf{w_j}, h_i)}{Q_i(h_i)}) \\ &\equiv \mathcal{F}(\boldsymbol{\Theta_w}, \mathbf{Q}) \end{aligned} \quad (13)$$

Thus $\mathcal{L}(\boldsymbol{\Theta_w})$ is lowerbounded by $\mathcal{F}(\boldsymbol{\Theta_w}, \mathbf{Q})$. The EM algorithm starts out with the initial values of Q, $\boldsymbol{\pi}$ and $\boldsymbol{\Theta_w}$ and iteratively optimizes them using co-ordinate wise ascent. In the E step, $\boldsymbol{\pi}$ and $\boldsymbol{\Theta_w}$ are held fixed and Q is optimized while in the M step $\boldsymbol{\pi}$ and $\boldsymbol{\Theta_w}$ are optimized holding Q fixed.

E-Step: Taking the derivative of $\mathcal{F}(\boldsymbol{\Theta_w}, \{Q_i(h_i)\}_{i=1}^M)$ w.r.t $Q_i(h_i)$ and setting the derivative to zero, we have

$$Q_i(h_i = j) = \frac{\pi_j \times P(\mathbf{x_i}|\mathbf{w_j}, h_i = j)}{\sum_{l=1}^{H}\pi_l \times P(\mathbf{x_i}|\mathbf{w_j}, h_i = j)} \quad (14)$$

using M Lagrange Multipliers, one for for each *document* to constrain Q to be a distribution over experts .

M-Step: Taking the derivative of $\mathcal{F}(\boldsymbol{\Theta_w}, \{Q_i(h_i)\}_{i=1}^M)$ w.r.t π_j holding $Q_i(h_i)$ fixed and setting the derivative to zero, we have

$$\pi_j = \frac{\sum_{i=1}^{M} Q_i(h_i = j)}{M} \quad (15)$$

again using Lagrange Multipliers for the experts.

Finally, taking the derivative of $\mathcal{F}(\boldsymbol{\Theta_w}, \{Q_i(h_i)\}_{i=1}^M)$ w.r.t $\mathbf{w_j}$ holding $Q_i(h_i)$ fixed, we have

$$\frac{\partial\mathcal{F}(\boldsymbol{\Theta_w}, \{Q_i(h_i)\}_{i=1}^M)}{\partial\mathbf{w_j}} = -\sum_{i=1}^{M} Q_i(h_i = j)\frac{\partial E_{i,j}}{\partial\mathbf{w_j}} + C \quad (16)$$

where C is some constant. Since the derivative depends on

$\mathbf{w_l}$'s, we use negative gradient descent to update $\mathbf{w_j}$ with

$$-\eta \triangle \mathbf{w_j} = \sum_{i=1}^{M} \eta\alpha Q_i(h_i = j)(f(\mathbf{x_i}^T\mathbf{w_j}))f(\mathbf{x_i}^T\mathbf{w_j})\times$$
$$(1 - f(\mathbf{x_i}^T\mathbf{w_j}))\mathbf{x_i} + \eta(\sum_{i=1}^{M} Q_i(h_i = j)\times$$
$$\sum_{l=1,l\neq j}^{H} \beta cos(\mathbf{w_j}^T\mathbf{w_l})sin(\mathbf{w_j}^T\mathbf{w_l})\mathbf{w_l}) \quad (17)$$

We choose the expert for datum i based on $argmax_i\mathbf{Q_i}(h_i)$ and assign the label t_i chosen be the expert for this datum. α was set to 0.03, β to 0.02 and η to 0.005 in all experiments.

5. RESULTS AND DISCUSSIONS

In this section, we present the results of our approach on a subset of the Spinner.com data [10] that was filtered for the recent US presidential election campaign blogs. For validating results, we chose 57 relevant manually annotated blog posts. 48 of them were labeled as pro-Obama and 9 of those were labeled as pro-McCain by the "objective" model. Using only KL divergence, the numbers of (proObama, proMcCain) blogs were found to be (24,29), (26,28), (26,31) and (39,18) by the baseline topic model, the "positive", the "negative" and the "objective" topic models respectively.

Table 1: Recall for the experimental test data set

Recall for Blogger Inclinations				
Blogger Inclination	CTM	positive-CTM	negative-CTM	objective-CTM
Obama	0.50	0.54	0.54	0.81
McCain	0.33	0.33	0.33	0.33

Some documents under some models were labeled as "cannotSay." The recall on this test set is given in table 1. This test set had been designed to be extremely skewed with respect to inclinations towards Obama and McCain and was done to ensure the effectiveness of the mixture of topic model "experts." Indeed the collaborative algorithm chose to label the documents output by the "objective" model. A similar manually labeled dataset with 48 pro-Obama and 30 pro-McCain blogs showed the same results with a negative deviation of 0.06 and 0.01 for the recall values for Obama and McCain respectively.

As we can observe from table 1, lack of sentiment leads to only 50% recall for candidate 1 which was Obama from the base model. However, when we weigh the base model with the probabilities of the positive, negative and objective sentiments for topics, we find the recall for candidate 1 (Obama) sharply rises for the "objective" topic model. This seems intuitive that the choice of words under the themes in both the speeches are negative and positive to the same extent and hence weigh the themes more or less equally. However, from the recent presidential election debates, we know that Obama used more specific and constructive i.e. objective words. Hence the themes discussed in blogs that maintain an objective view are the blogs that tend to support the themes of Obama more than the negativity prevalent in McCain's speeches.

In table 2, we observe the effect of this theme matching between those of the blog and those in the candidate speeches. In our method, the topics in *each blog document* is *inferred* and for such a blog document, the theme distribution of each word is compared to that for an overlapping word in the ground truth speeches.

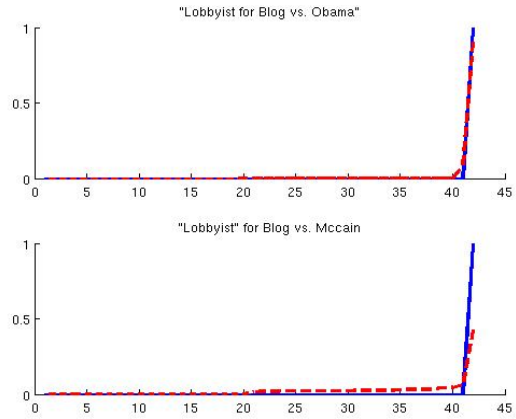

Figure 2: Word topic distributions for "lobbyist" for a blog vs. Obama and McCain speeches

Table 2: Positive Theme orientation between a blog and candidate speeches for the "Objective" Topic Model

Topic similarities			
Dataset ↓	term in blog	topic in candidate speech	Few most probable stemmed words under the topic
Obama	lobbyist	4	promis america work american time countri chang live care tax peopl famili democrat economi job
McCain	lobbyist	23	campaign outrag time press polit voter point charg convent discuss rate emot shift attack
Blog#26	lobbyist	12	wall street polici campaign peopl secur time econom blogger social regul crisi
BlogText	.... as Mccain tries to pander to his lobbyist pals and the Republican pro-gun base but wanders into the War On Some Terror minefield by mistake.		

Table 2 shows two words from a blog that was labeled as pro-Obama. A few most probable words are shown for the corresponding maximally activated topic in the candidate speech collection. The topic responsible for activation of the term "lobbyist" in the blog document is similar to topic 4 in Obama's speeches where "lobbyists" were mentioned as one of the reasons for economic crisis and a need for change. McCain's topic for "lobbyist" contained words perhaps meant to blame the other party. Note that because of the unimportance of the topic ID assigned by the algorithm, we need to sort the term over topic probability density vector before computing the divergence. This density vector of the word "lobbyist" is shown in Figure 2 (blue for blog and red dashes for candidates). The X axis are the sorted topic numbers and the Y axis represent the density. Clearly, the graph for the blog is similar in "shape" to that for Obama for the word "lobbyist".

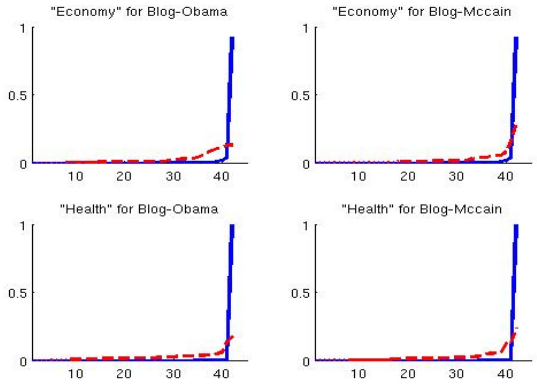

Figure 3: Word topic distributions for "economy" and "health" for a blog vs. Obama and McCain speeches

The reason why the KL divergence alone is not a deciding measure to indicate blogger orientation towards a particular candidate becomes evident when we look at Figure 3 and Table 3. The excerpt of the blog text in Table 3 indicates that the blogger is supporting Obama's ideas using his personal views. However, the blog post was determined to be oriented more towards McCain due to the majority of blog word-topic distributions having "slightly" lower KL divergence w.r.t McCain's ground truth speeches. If we look carefully at the red graphs for Obama, the word "economy" was not only emphasized in only one theme related to the economy but also in other themes where economy was playing a major role. The use of the word "health" was mostly confined to one theme only and this was same for both the ground truth speech sets.

Since the goal was the determination of the reason for political orientation of a blogger w.r.t topics shared and was not posed as a typical classification problem based on "for-against" annotation of the blogs, we observe that a mojority voting of lower KL divergences for the overlapping words of the blogs and ground truth speeches is not robust enough.

Table 3: "Misclassification" due to nearly equal KL divergence between Blogs and candidate speeches for the Objective Model

Topic similarities			
Blog vs. ↓	term in blog	topic in candidate speech	Few most probable stemmed words under the topic
Obama	ecomony	0	energi oil invest effici feder nation percent technolog fuel build advanc develop
	health	18	plan make cost health american care presid system famili requir
McCain	ecomony	24	plan tax econom account save job economi stock propos measur
	health	41	bush administr republican support care state issu campaign peopl forc talk health
BlogText	Bush-McCain policies have ruined the US economy says this new Obama ad, and now MCain wants to do the same to our health care		

Indeed, precisely because of the phenomenon decribed above in Obama's speeches, mislabeling of blogs as pro-Obama or pro-McCain was common. However, the candidate topics in tables 2 and 3 are indicator enough as to why a blogger is more oriented towards a particular candidate.

The other and more subtle effect of sarcasms couldn't be neglected either. Some articles have a vocabulary that is dominated by all terms related to the promises made by one candidate, but ends with a sentence that changes the overall tone of the article. Some articles are humor based where all the policies made by a candidate is debated using sarcasm or jokes. This is a good reason why the word topic probabilities needed to be weighted by the topic sentiment.

Consider the text - *"Paris Hilton has responded to John McCain for including her in one of his recent campaign ads. To prevent any confusion, please note that, although she does make more sense than John McCain, ... McCain policy loses all substance once you get beyond his attacks on Obama. Paris did make some points but they were fundamentally flawed."* Firstly, words like Paris and Hilton occurred on some unbiased news reports for McCain only. Even if this isn't the case, topic inference-wise many proMcCain blogs got associated with Obama's themes because in the real-life blog data it was observed that such bloggers were quoting themes mostly activated by words used by Obama or were used in the reports about him except that the bloggers were using them for sarcastic criticisms.

It is important to note here that with regard to the noisy data issue, readers may be confused as to what is noisy in this problem setting. Blog data may inherently be noisy in terms of irrelevant tokens and spelling variations and can be cleaned quite effectively using standard information retrieval and other tools - a issue not very relevant. What is considered noisy here is that many blogs specially politico-social blogs are written with subtelty in "meaning" of the text - what appears on a first reading is really not what is implied. To this extent, we considered the presence of such subtle implication shifts through author tonality as noise. The presence of these shifts imposes additional information load on part of the readers to understand the true intensions which are revealed after thorough conceptual filtering. Note that this view of noise is not the same as splogs where the intents are clear but not at all relevant to a particular blog topic discussion. Any phenomenon that increases the information load on part of the end user to understand the underlying structure may be attributed to noise. An example in the image domain may comprise of the fact that addition of salt and pepper noise greatly degrades the visual quality of an image from easy understanding.

A very recent work by [13] shows one way of jointly modeling sentiments with topics using sentiment coverages. While it may be useful to explore this alternative, however, we believe it might be more useful to jointly model the (blog collection, candidate ground-truth) pair with sentiment and semantic contents and then compare document theme similarities. A similar line of thinking has been explored in [15] where a joint modeling is performed to predict blog comments from the head blog post.

6. CONCLUSION

To our knowledge, there has been no prior research on this exact problem. Detecting sentiments in blogs is quite a different phenomenon where the goal is to predict a blogger is showing a positive, negative or neutral sentiment. We believe that to properly address this problem, three things need to be done. Firstly, a ground truth collection should be jointly modeled with the blog collection taking into account sentiment and "tonality". Secondly, the inference under such a joint model should address the issue of predicting new blog posts given (ground-truth, training-blog) pairs of documents. Finally, after modeling every possible (ground-truth-collection, blog-collection) pair for each of the candidates, the likelihood of the blogs under the different collections should be compared to determine a probabilistic classification of the blogger towards a particular candidate.

In this paper we have discussed a topic model based approach to identify blogger orientation towards a particular election candidate. Topic models provide an effective way to explain why a blogger is biased towards a condidate for the themes he speaks of rather than his charisma. However, the use of KL divergence alone is not an effective form of *"classification"* because subtle numerical differences in divergences in this unsupervised setting can lead to a mis-labeling of sarcastic blogs. Detecting sarcasm in text is indeed very hard and remains an open problem.

7. REFERENCES

[1] Lada Adamic and Natalie Glace. The political blogosphere and the 2004 u.s. election: divided they blog. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, pages 36–43, 2005.

[2] David Blei and John Lafferty. Correlated topic models. In *Advances in Neural Information Processing Systems*, volume 18, 2005.

[3] David Blei and Jon McAuliffe. Supervised topic models. In *Advances in Neural Information Processing Systems*, volume 20, 2008.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] Eric Breck, Yejin Choi, and Claire Cardie. Identifying expressions of opinion in context. In *Twentieth International Joint Conference on Artificial Intelligence*, 2007.

[6] Yejin Choi and Claire Cardie. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2008.

[7] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[8] Kathleen Durant and Michael Smith. Mining sentiment classification from political web logs. In *Proceedings of Workshop on Web Mining and Web Usage Analysis of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (WebKDD-2006)*, 2006.

[9] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC*, pages 417–422, 2006.

[10] ICWSM. Icwsm 2009 spinn3r dataset. In *Proceedings of the Third International Conference on Weblogs and Social Media (ICWSM 2009)*, San Jose, CA, May 2009.

[11] Michael I. Jordan. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6:181–214, 1994.

[12] Frank Lin and William W. Cohen. The multirank bootstrap algorithm: Semi-supervised political blog classification and ranking using semi-supervised link classification. In *ICWSM'08 Poster*, 2008.

[13] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and Chengxiang Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *In Proc. of the 16th Int. Conference on World Wide Web*, pages 171–180, 2007.

[14] Tony Mullen and Robert Malouf. A preliminary investigation into sentiment analysis of informal political discourse. In *Proceedings of the AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.

[15] Tae Yano, William W. Cohen, and Noah A. Smith. Predicting response to political blog posts with topic models. In *Proceedings of NAACL HLT*, page TBD, 2009.