Mining Contentions from Discussions and Debates

Arjun Mukherjee and Bing Liu
Department of Computer Science
University of Illinois at Chicago
arjun4787@gmail.com, liub@cs.uic.edu

ABSTRACT
Social media has become a major source of information for many applications. Numerous techniques have been proposed to analyze network structures and text contents. In this paper, we focus on fine-grained mining of contentions in discussion/debate forums. Contentions are perhaps the most important feature of forums that discuss social, political and religious issues. Our goal is to discover contention and agreement indicator expressions, and contention points or topics both at the discussion collection level and also at each individual post level. To the best of our knowledge, limited work has been done on such detailed analysis. This paper proposes three models to solve the problem, which not only model both contention/agreement expressions and discussion topics, but also, more importantly, model the intrinsic nature of discussions/debates, i.e., interactions among discussants or debaters and topic sharing among posts through quoting and replying relations. Evaluation results using real-life discussion/debate posts from several domains demonstrate the effectiveness of the proposed models.

Categories and Subject Descriptors
H.2.8 [Database Management]: Database Applications – *Data mining*; H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing.

General Terms
Algorithms, Experimentation

Keywords
Contention Analysis, Debates, Discussions, Social Media

1. INTRODUCTION
Social media such as reviews, blogs, comments, discussions and debates contain valuable information that can be used for many applications. The essence of such media is that it enables people from anywhere in the world to express their views and to discuss any issue of interest in online communities. A large part of such discussions is about social, political and religious issues. On such issues, there are often heated discussions/debates, i.e., people argue and agree or disagree with one another. In this paper, we model this form of interactive social media. Given a set of discussion/debate posts, we aim to perform the following tasks:

1. Discover expressions that people often use to express contention (e.g., "*I disagree*", "*you make no sense*") and agreement (e.g., "*I agree*", "*I think you're right*"). We collectively call them *CA-expressions*.
2. Determine contentious topics. First discover discussion topics in the whole collection, and then for each contentious post, discover the contention points (or topics).

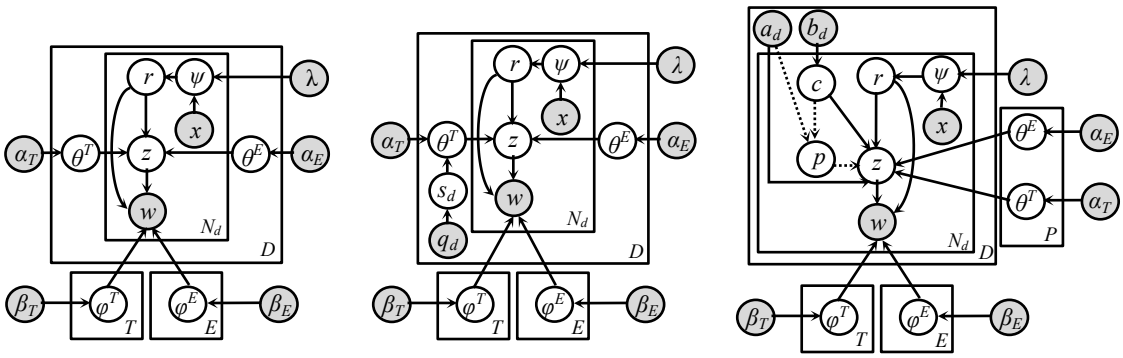
KDD'12, August 12–16, 2012, Beijing, China.

Although there is a large body of literature on social media analysis such as social network analysis [14], sentiment analysis [27, 34], and grouping people into different camps [1, 32, 40, 42], to the best of our knowledge, limited research has been done on the fine-grained analysis of discussion/debate forums as proposed in this paper. This problem is important because a large part of social media is about discussions/debates of contentious issues, and discovering such issues is useful for many applications. For example, in a political election, contentious issues often separate voters into different camps and determine their political orientations. It is thus important for political candidates to know such issues. For contentious social topics, it is crucial for government agencies to be aware of them so that they can address the problems. Even for consumer products/services, contentions about them can be used to identify different types of customers and to make effective marketing and business decisions.

In this paper, we use statistical modeling to perform the aforementioned tasks. Three new models are proposed. The first model, called JTE (Joint Topic-Expression model), jointly models both discussion topics and CA-expressions. It provides a general framework for discovering discussion topics and CA-expressions simultaneously. Its generative process separates topics and CA-expressions by using maximum-entropy priors to guide the separation. However, this model does not consider a key characteristic of discussions/debates, i.e., authors quote or mention the claims/views of other authors and express contention or agreement on those claims/views. That is, there are interactions among authors and topics through the reply-to relation, which is a salient feature of discussion/debate forums. We then extend the JTE model and propose two novel and more advanced models JTE-R and JTE-P which model the interactions of authors and topics in two different ways, based on reply-to relations and author-pair structures respectively.

Works related to ours are quite different both in application and in modeling. On application, the closely related work to ours is the finding of different camps of people in discussions/debates [1, 32, 40, 42]. This thread of research, however, does not discover CA-expressions or contention points, which are the objectives of this work. From a modeling point of view, our work is related to topic modeling in general and joint modeling of topics and certain other information in particular. Topic models are a principled way of mining topics from large text collections. There have been many extensions [5, 7, 13, 29, 35, 43] to the initial models, LDA (Latent Dirichlet Allocation) [4] and pLSA (Probabilistic latent semantic analysis) [20]. However, these models mine only topics, which are insufficient for our problem. In recent years, researchers have also proposed joint models of topics and sentiments [21, 26, 30, 47]. Our JTE model is related to these joint models. However, these models treat documents/posts independently, which fail to capture author and topic interactions in discussions/debates, i.e., authors reply to and quote each other's claims/views and express contentions or agreements. Due to such interactions, posts are clearly not independent of one another. The proposed JTE-R and JTE-P models capture such interactions. The detailed comparison with these models and other related work appears in §5.

The proposed models are evaluated both qualitatively and quantitatively using a large number of real-life discussion/debate posts from four domains. Experimental results show that the two



(a) JTE Model (b) JTE-R Model (c) JTE-P Model

Figure 1: Graphical Models: Plate notations of JTE, JTE-R, JTE-P models. Shaded and unshaded variables indicate observed and latent variables respectively. Note that the pair variable p in the JTE-P model is introduced for derivational convenience and thus its causalities are shown by dotted arrows.

advanced models (JTE-R and JTE-P) outperform the base JTE model significantly, which indicate that the proposed interaction modeling is effective. Experiments using statistical metrics such as perplexity and KL-Divergence also demonstrate that the interaction models fit the discussion/debate data better and can discover more distinctive topics and CA-expressions.

2. JTE MODEL

This section presents the proposed JTE (Joint Topic Expression) model which lays the ground work for jointly modeling topics and CA-expressions. The JTE model belongs to the family of generative models for text where words and phrases (*n*-grams) are viewed as random variables, and a document is viewed as a bag of *n*-grams and each *n*-gram (word/phrase) takes one value from a predefined vocabulary. We use up to 4-grams, i.e., n = 1, 2, 3, 4, in this work. Note that topics in most topic models like LDA are usually unigram distributions over words and assume words to be exchangeable at the word level. Arguably, this offers a great computational advantage over more complex models taking word order into account for discovering significant *n*-grams [44, 45]. Yet there exist other works which try to *post-process* discovered topical unigrams to form multi-word phrases using relevance [46] and likelihood [6] scores. However, our goal in this work is to enhance the expressiveness of our JTE model (rather than modeling *n*-gram word order) by considering *n*-grams and preserving the advantages of exchangeable modeling. Thus, we consider both words and phrases as our vocabulary (more details in §4.1). For notational convenience, from now on we use *terms* to denote both *words* (unigrams) and *phrases* (*n*-grams). Since we are dealing with large corpora, for computational reasons, we only consider terms which appeared at least 30 times in the corpus[1]. We denote the entries in our vocabulary by $v_{1\ldots V}$, where V is the number of unique terms in the vocabulary. The entire corpus (document collection) of study is comprised of $d_{1\ldots D}$ documents. A document (e.g., discussion post) d is represented as a vector of terms $\boldsymbol{w_d}$ with N_d entries. W is the set of all observed terms in the corpus with cardinality, $|W| = \sum_d N_d$. The JTE model is motivated by the joint occurrence of CA-expression types (*contention* and *agreement*) and topics in discussion posts. A typical discussion/debate post mentions a few topics (using semantically related topical terms) and expresses some viewpoints with one or more CA-expression types (using semantically related contention and/or agreement expressions). This observation motivates the generative process of our model where documents (posts) are represented as random mixtures of latent topics and CA-expression types. Each topic or CA-expression type is characterized by a distribution over terms. Assume we have t = 1,…, T topics and e = 1,…, E expression types in our corpus. Note that in our case of discussion/debate forums, based on reading various posts, we hypothesize that E = 2 as in such forums, we mostly find two expression types: contention and agreement (which we also statistically validate in §4.4.1). However, the proposed JTE and other models are general and can be used with any number of expression types. Let $\psi_{d,j}$ denote the probability of $w_{d,j}$ being a topical term with $r_{d,j} \in \{\hat{t}, \hat{e}\}$ denoting the binary indicator variable (topic or CA-expression) for the j^{th} term of d, $w_{d,j}$. $z_{d,j}$ denotes the appropriate topic or CA-expression type index to which $w_{d,j}$ belongs. We parameterize multinomials over topics using a matrix $\Theta^T_{D\times T}$ whose elements $\theta^T_{d,t}$ signify the probability of document d exhibiting topic t. For simplicity of notation, we will drop the latter subscript (t in this case) when convenient and use θ^T_d to stand for the d^{th} row of Θ^T. Similarly, we define multinomials over CA-expression types using a matrix $\Theta^E_{D\times E}$. The multinomials over terms associated with each topic are parameterized by a matrix $\Phi^T_{T\times V}$, whose elements $\varphi^T_{t,v}$ denote the probability of generating v from topic t. Likewise, multinomials over terms associated with each CA-expression type are parameterized by a matrix $\Phi^E_{E\times V}$. We now define the generative process of JTE (see Figure 1(a) for plate notation).

1. For each CA-expression type e, draw $\varphi^E_e \sim Dir(\beta_E)$
2. For each topic t, draw $\varphi^T_t \sim Dir(\beta_T)$
3. For each forum discussion post $d \in \{1 \ldots D\}$:
   i. Draw $\theta^E_d \sim Dir(\alpha_E)$
   ii. Draw $\theta^T_d \sim Dir(\alpha_T)$
   iii. For each term $w_{d,j}$, $j \in \{1 \ldots N_d\}$:
      a. Set $\psi_{d,j} \leftarrow MaxEnt(\overrightarrow{x_{d,j}};\ \lambda)$
      b. Draw $r_{d,j} \sim Bernoulli(\psi_{d,j})$
      c. if $(r_{d,j} = \hat{e})$ // $w_{d,j}$ is a CA-expression term
            Draw $z_{d,j} \sim Mult(\theta^E_d)$
         else // $r_{d,j} = \hat{t}$, $w_{d,j}$ is a topical term
            Draw $z_{d,j} \sim Mult(\theta^T_d)$
      d. Emit $w_{d,j} \sim Mult(\varphi^{r_{d,j}}_{z_{d,j}})$

We use Maximum Entropy (Max-Ent) model to set $\psi_{d,j}$. The Max-Ent parameters can be learned from a small number of labeled topical and CA-expression terms which can serve as good priors. The idea is motivated by the following observation: topical and CA-expression terms usually play different syntactic roles in a sentence. Topical terms (e.g. "U.S. senate", "sea level", "marriage", "income tax") tend to be noun and noun phrases while CA-expression terms ("I refute", "how can you say", "probably agree") usually contain pronouns, verbs, wh-determiners, and modals. In order to utilize the part-of-speech (POS) tag information, we place $\psi_{d,j}$ (the prior over the indicator variable $r_{d,j}$) in the word plate (see Figure 1 (a)) and draw it from a Max-Ent model conditioned on the observed feature vector $\overrightarrow{x_{d,j}}$ associated with $w_{d,j}$ and the learned Max-Ent parameters λ (see §

[1] This is reasonable as our corpus contains about 100,000 documents (see §4). It is unlikely for a term with frequency less than 30 to show up as a top topical or CA-expression term.

4.1). $\overrightarrow{x_{d,j}}$ can encode arbitrary contextual features that may be discriminative. In this work, we encode both lexical and POS features of the previous, current and next POS tags/lexemes of the term $w_{d,j}$. Specifically, the feature vector, $\overrightarrow{x_{d,j}} = [POS_{w_{d,j}-1}, POS_{w_{d,j}}, POS_{w_{d,j}+1}, w_{d,j}-1, w_{d,j}, w_{d,j}+1]$. For phrasal terms ($n$-grams), all POS tags and lexemes of $w_{d,j}$ are considered as features. To learn the JTE model from data, exact inference is not possible. We thus resort to approximate inference using collapsed Gibbs sampling [18]. We first derive the joint distribution below and then the Gibbs sampler.

To derive the joint distribution, we factor the joint according to the conditional distributions (causalities) governed by the Bayesian network of the proposed generative model.

$$P(W,Z,R) = P(W|Z,R) \times P(Z|R) \times P(R) \quad (1)$$

Since we employ a collapsed Gibbs sampler, we integrate out θ and φ and obtain the joint as follows.

$$P(W,Z,R) = \left[\prod_{t=1}^{T} \frac{B(n_t^{TV}+\beta_T)}{B(\beta_T)} \times \prod_{e=1}^{E} \frac{B(n_e^{EV}+\beta_E)}{B(\beta_E)}\right] \times \left[\prod_{d=1}^{D}\left(\frac{B(n_d^{DT}+\alpha_T)}{B(\alpha_T)} \times \frac{B(n_d^{DE}+\alpha_E)}{B(\alpha_E)}\right)\right] \times \left[\prod_{d=1}^{D}\prod_{j=1}^{N_d} p(r_{d,j}|\psi_{d,j})\right] \quad (2)$$

where $p(r_{d,j}|\psi_{d,j}) = (\psi_{d,j,\hat{t}})^{u}(\psi_{d,j,\hat{e}})^{1-u}$; $u = \begin{cases}1, r_{d,j} = \hat{t}\\ 0, r_{d,j} = \hat{e}\end{cases}$ and the outcome probabilities of the Max-Ent model are given by:

$$\psi_{d,j,\hat{t}} = p(y = \hat{t}|x_{d,j}); \psi_{d,j,\hat{e}} = p(y = \hat{e}|x_{d,j});$$

$$p(y|x_{d,j}) = \frac{\exp(\sum_{i=1}^{n}\lambda_i f_i(x_{d,j},y))}{\sum_{y\in\{\hat{t},\hat{e}\}}\exp(\sum_{i=1}^{n}\lambda_i f_i(x_{d,j},y))}$$

$\lambda_{1\ldots n}$ are the parameters of the learned Max-Ent model corresponding to the n binary feature functions $f_{1\ldots n}$ from Max-Ent. $n_{t,v}^{TV}$ and $n_{e,v}^{EV}$ denote the number of times term v was assigned to topic t and expression type e respectively. $B(\cdot)$ is the multinomial Beta function $B(\vec{x}) = \frac{\prod_{i=1}^{\dim(\vec{x})}\Gamma(x_i)}{\Gamma(\sum_{i=1}^{\dim(\vec{x})}x_i)}$. $n_{d,t}^{DT}$ and $n_{d,e}^{DE}$ denote the number of terms in document d that were assigned to topic t and CA-expression type e respectively. n_t^{TV}, n_e^{EV}, n_d^{DT}, and n_d^{DE} denote the corresponding row vectors.

We use Gibbs sampling for posterior inference. Gibbs sampling is a form of Markov chain Monte Carlo (MCMC) method where a Markov chain is constructed to have a particular stationary distribution. In our case, we want to construct a Markov chain which converges to the posterior distribution over R and Z conditioned on the observed data. We only need to sample z and r as we use collapsed Gibbs sampling and the dependencies of θ and φ have already been integrated out analytically in the joint. Denoting the random variables $\{w,z,r\}$ by singular subscripts $\{w_k,z_k,r_k\}$, $k_{1\ldots K}$, $K = \sum_d N_d$, a single iteration consists of performing the following sampling:

$$p(z_k = t, r_k = \hat{t}|Z_{\neg k}, W_{\neg k}, R_{\neg k}, w_k = v) \propto \frac{n_{d,t\ \neg k}^{DT} + \alpha_T}{n_{d,(\cdot)\ \neg k}^{DT} + T\alpha_T} \frac{n_{t,v\ \neg k}^{TV}+\beta_T}{n_{t,(\cdot)\ \neg k}^{TV}+V\beta_T} \times \frac{\exp(\sum_{i=1}^{n}\lambda_i f_i(x_{d,j},\hat{t}))}{\sum_{y\in\{\hat{t},\hat{e}\}}\exp(\sum_{i=1}^{n}\lambda_i f_i(x_{d,j},y))} \quad (3)$$

$$p(z_k = e, r_k = \hat{e}|Z_{\neg k}, W_{\neg k}, R_{\neg k}, w_k = v) \propto \frac{n_{d,e\ \neg k}^{DE} + \alpha_E}{n_{d,(\cdot)\ \neg k}^{DE} + E\alpha_E} \times \frac{n_{e,v\ \neg k}^{EV}+\beta_E}{n_{e,(\cdot)\ \neg k}^{EV}+V\beta_E} \times \frac{\exp(\sum_{i=1}^{n}\lambda_i f_i(x_{d,j},\hat{e}))}{\sum_{y\in\{\hat{t},\hat{e}\}}\exp(\sum_{i=1}^{n}\lambda_i f_i(x_{d,j},y))} \quad (4)$$

where $k = (d,j)$ denotes the j^{th} term of document d and the subscript $\neg k$ denotes assignments excluding the term at k. Omission of a latter index denoted by $(\cdot)$ represents the marginalized sum over the latter index. The conditional probabilities in (3) and (4) were derived by applying the chain rule on the joint distribution. We employ a blocked sampler where we sample r and z jointly, as this improves convergence and reduces autocorrelation of the Gibbs sampler [37].

3. JTE-R and JTE-P MODELS

We now augment the JTE model to encode two kinds of interactions (topic and author) in discussion/debate forums.

3.1 JTE-R: Encoding Reply Relations

We first improve JTE by encoding the reply-to relations as authors usually reply to each other's viewpoints by explicitly mentioning the user name using @name, and/or by quoting others' posts. For easy presentation, we refer both cases by *quoting* from now on. Considering the reply-to relation, we call the new model JTE-R (Figure 1 (b)). This model is based on the following observation:

Observation: Whenever a post d replies to the viewpoints in some other posts by quoting them, d and the posts quoted by d should have similar topic distributions.

This observation indicates that the JTE-R model needs to depart from typical topic models where there is usually no topical interaction among documents, i.e., documents are treated as being independent of one another. Let q_d be the set of posts quoted by post d. Clearly, q_d is observed[2]. In order to encode this "reply-to" relation into our model, the key challenge is to somehow constrain the topic distribution of d, θ_d^T to be similar to the topic distributions of posts in q_d. Specifically, it is how to constrain θ_d^T to be similar to $\theta_{\acute{d}}^T$, where $\acute{d} \in q_d$ (i.e., constraining topic assignments to documents) during inference while the topic distributions of both θ_d^T and $\theta_{\acute{d}}^T$, $\acute{d} \in q_d$ are latent and unknown *apriori*. Note that this situation is very different from that in [2] where it constrains word assignments to topics *apriori* knowing that some words are semantically related and probably should belong to the same topic. To solve our problem, we propose a novel solution, which exploits the following salient features of the Dirichlet distribution:

1. Since $\theta_d^T \sim Dir(\alpha_T)$, we have $\sum_t \theta_{d,t}^T = 1$. Thus, it suffices that θ_d^T can act as a base measure for Dirichlet distributions of the same order.
2. Also, the expected probability mass associated with each dimension of the Dirichlet distribution is proportional to the corresponding component of its base measure[3].

Thus, to constrain a post d's topic distribution to be similar to the posts whom it replies/quotes (i.e. posts in q_d), we now need functional base measures as it is the base measure that governs the expected mass associated with each topical dimension in θ_d^T. One way to employ functional base measures is to draw $\theta_d^T \sim Dir(\alpha_T \boldsymbol{s_d})$, where $\boldsymbol{s_d} = \sum_{d'\in q_d}\theta_{d'}^T/|q_d|$ (the expected topical distribution of posts in q_d). For posts which do not quote any other post, we simply draw $\theta_d^T \sim Dir(\alpha_T)$. For such a topic model with functional Dirichlet base measures, the sampling distribution is more complicated. Specifically, the document-topic distribution, θ_d^T is no longer a simple predictive distribution, i.e., when sampling z_n^d, the implication of each quoted document related to d by reply-to relations and their topic assignments must be considered because the sampling distribution for z_n^d in document d must consider its effect on the joint probability of the entire model. Unfortunately, this too can be computationally expensive for large corpora (like ours). To circumvent this issue, we can hierarchically sample documents based on reply-to relation network using sequential Monte Carlo [9], or approximate the true Gibbs sampling distribution by updating the original smoothing parameter (α_T) to reflect the expected topic distributions of quoted documents ($s_{d,t}\alpha_T$), where $s_{d,t}$ is the t^{th} component of the base measure, s_d which is computed at runtime during sampling. We take the latter approach (see Eq. (5)). Experiments show that this

[2] We crawled the ids of posts quoted by each post.

[3] Taking moments on $(X_1 \ldots X_n) \sim Dir(\alpha_1 \ldots \alpha_n)$, we get $E[X_i] = \frac{\alpha_i}{\sum\alpha_i}$. Thus, $E[X_i] \propto \alpha_i$

approximation performs well empirically. The approximate Gibbs distribution for JTE-R while sampling $z_n^d = t$ is given by:

$$p(z_k = t, r_k = \hat{t} | Z_{\neg k}, W_{\neg k}, R_{\neg k}, w_k = v) \propto \frac{n_{d,t\ \neg k}^{DT} + s_{d,t}\alpha_T}{\sum_{t=1}^{T}\left(n_{d,t\ \neg k}^{DT} + s_{d,t}\alpha_T\right)} \frac{n_{t,v\ \neg k}^{TV} + \beta_T}{n_{t,(\cdot)\ \neg k}^{TV} + V\beta_T} \times \frac{\exp(\sum_{i=1}^{n}\lambda_i f_i(x_{d,j}, \hat{t}))}{\sum_{y\in\{\hat{t},\hat{e}\}}\exp(\sum_{i=1}^{n}\lambda_i f_i(x_{d,j}, y))} \quad (5)$$

3.2 JTE-P: Encoding Pair Structures

JTE-R builds over JTE by encoding reply-to relations to constrain a post to have similar topic distributions to those it quotes. An alternative strategy is to make θ^T and θ^E author-pair specific. The idea is motivated by the following observation.

Observation: When authors reply to others' viewpoints (by @name or quoting other authors' posts), they typically direct their own topical viewpoints with contentious or agreeing expressions to those authors. Such exchanges can go back and forth between pairs of authors. The discussion topics and CA-expressions emitted are thus caused by the author-pairs' topical interests and their nature of interactions.

Let a_d be the author of a post d, and $b_d = [b_{1\ldots n}]$ be the list of *target authors* (we will also call them *targets* for short) to whom a_d replies to or quotes in d. The pairs of the form $p = (a_d, c)$, $c \in b_d$ essentially shapes both the topics and CA-expressions emitted in d as contention or agreement on topical viewpoints are almost always directed towards certain target authors. For example, if c claims something, a_d quotes the claim in his post d and then contends/agrees by emitting CA-expressions like "you have no clue", "yes, I agree", "I don't think," etc. Clearly, this pair structure is a crucial feature of discussions/debate forums. Each pair has its unique and shared topical interests and interaction nature (contention or agreement). Thus, it is appropriate to condition θ^T and θ^E over author-pairs. We will see in §4.4.1 that this model fits the discussion data better. Standard topic models do not consider this key piece of information. Although there are extensions to consider authors [37], persona [31] and interest [24], none of them are suitable for considering the pair structure.

We extend the JTE model to incorporate the pair structure. We call the new model JTE-P, which conditions the multinomial distributions over topics and CA-expression types (θ^T, θ^E) on authors and targets as pairs rather than on documents as in JTE and JTE-R. In its generative process, for each post, the author a_d and the set of targets b_d are observed. To generate each term $w_{d,j}$, a target, $c \sim Uni(b_d)$, is chosen at uniform from b_d forming a pair $p = (a_d, c)$. Then, depending on the switch variable $r_{d,j}$, a topic or an expression type index z is chosen from a multinomial over topic distribution θ_p^T or CA-expression type distribution θ_p^E, where the subscript p denotes the fact that the distributions are specific to the author-target pair p which shape topics and CA-expressions. Finally, the term is emitted by sampling from topic or CA-expression specific multinomial distribution $\varphi_{z_{d,j}}^{r_{d,j}}$.

The graphical model in plate notation corresponding to the above process is shown in Figure 1 (c). Clearly, in JTE-P, the discovery of topics and CA-expressions are guided by the pair structure of reply-to relations in which the collection of posts was generated. For posterior inference, we again use Gibbs sampling. Note that as a_d is observed, sampling c is equivalent to sampling the pair $p = (a_d, c)$. Its Gibbs sampler is given by:

$$p(z_k = t, p_k = p, r_k = \hat{t} | \ldots, w_k = v) \propto \frac{1}{|b_d|} \times \frac{n_{p,t\ \neg k}^{PT} + \alpha_T}{n_{p,(\cdot)\ \neg k}^{PT} + T\alpha_T} \times \frac{n_{t,v\ \neg k}^{TV} + \beta_T}{n_{t,(\cdot)\ \neg k}^{TV} + V\beta_T} \times \frac{\exp(\sum_{i=1}^{n}\lambda_i f_i(x_{d,j}, \hat{t}))}{\sum_{y\in\{\hat{t},\hat{e}\}}\exp(\sum_{i=1}^{n}\lambda_i f_i(x_{d,j}, y))} \quad (6)$$

$$p(z_k = e, p_k = p, r_k = \hat{e} | \ldots, w_k = v) \propto \frac{1}{|b_d|} \times \frac{n_{p,e\ \neg k}^{PE} + \alpha_E}{n_{p,(\cdot)\ \neg k}^{PE} + E\alpha_E} \times \frac{n_{e,v\ \neg k}^{EV} + \beta_E}{n_{e,(\cdot)\ \neg k}^{EV} + V\beta_E} \times \frac{\exp(\sum_{i=1}^{n}\lambda_i f_i(x_{d,j}, \hat{e}))}{\sum_{y\in\{\hat{t},\hat{e}\}}\exp(\sum_{i=1}^{n}\lambda_i f_i(x_{d,j}, y))} \quad (7)$$

where $n_{p,t}^{PT}$ and $n_{p,e}^{PE}$ denote the number of times the pair p was assigned to topic t and expression type e respectively. As JTE-P assumes that each pair has a specific topic and expression distribution, we see that Eq. (6, 7) shares topics and expression types across pairs. It is also worthwhile to note that given A authors, there are $\binom{A}{2}$ possible pairs. However, the actual number of pairs (i.e., where the authors have communicated at least once) is much less than $\binom{A}{2}$. Our experimental data consists of 1824 authors and 7684 actual pairs. Hence we are only modeling 7684 pairs instead of $\binom{1824}{2} \approx 4$ million pairs.[4]

4. EXPERIMENTS

We now evaluate the proposed models and compare with baselines. We first qualitatively show the CA-expressions and topics discovered by the models. We then evaluate the models quantitatively based on the two objectives of this work:

i) Discovering contention and agreement expressions (or CA-expressions).
ii) Finding contention points or topics in each contentious post.

Experiments are also conducted on statistical metrics such as perplexity and KL-Divergence. They show that the interaction models (JTE-R and JTE-P) fit the discussion/debate data better and find more distinctive topics and CA-expressions than JTE.

For our experiments, we used debate/discussion posts from Volconvo[5]. The forum is divided into various domains: Politics, Religion, Society, Science, etc. Each domain consists of multiple threads. Each thread has a set of posts. For each post, we extracted the post id, author, time, domain, ids of all posts to which it replies/quotes, and the post content. In all, we extracted 26137, 34986, 22354 and 16525 posts from Politics, Religion, Society and Science domains respectively. Our final data consists of 5978357 tokens, 1824 authors with an average of 346 words per post, and 7684 author-target pairs.

4.1 Topic and CA-expression Discovery

To set the background for our quantitative evaluation of the two tasks in the next two sub-sections, we first show the topics and the CA-expressions discovered by our models, and also compare them with topics found by LDA [4] and its variant SLDA (sentence-LDA) [21]. We choose LDA as it is the best-known topic model. We use SLDA as it constrains words in a sentence to be generated from a single topic [21]. Since CA-expressions may appear with topics in the same sentence, we want to see how it performs in mining topics and CA-expressions although SLDA is unable to separate topical terms and CA-expressions.

For all our experiments here and the subsequent ones, the hyper-parameters for LDA and SLDA were set to the heuristic values $\alpha = 50/T$, $\beta = 0.1$ as suggested in [18]. Similarly, for the proposed models, we set $\alpha_T = 50/T$, $\alpha_E = 50/E$, $\beta_T = \beta_E = 0.1$. To learn the Max-Ent parameters λ, we randomly sampled 500 terms from our corpus appearing at least 50 times[6] and labeled them as topical (372) or CA-expressions (128) and used the corresponding feature vector of each term (in the context of posts where it

[4] Note that in modeling pair interactions, JTE-P conditions topic and CA-expression emission over debating pairs. The model could be made further fine-grained by modeling topic specific interaction of authors by incorporating some ideas in [16, 37] While this approach is richer, it calls for more complex inference. We will study these possibilities in the future. We thank one anonymous reviewer for this suggestion.

[5] http://www.volconvo.com/forums/forum.php

[6] A minimum frequency count of 50 ensures that the training data is reasonably representative of the corpus.

(a) JTE

t_1: Spirituality	t_2: Burka/Veil	t_3: Homo-sexuality	t_4: Evolution of Life	t_5: 9/11 Attacks	t_6:Theism/Atheism	t_7: Global warming	t_8: Vegetarianism	t_9: IRS/Taxes	t_{10}: U.S. Politics
spirituality	burka	marriage	evolution	9/11	god	earth	animals	tax	vote
life	burqa	gay	species	september 11	belief	co2	meat	government	president
soul	suffering	couples	theory	terrorism	**existence**	warming	egg	irs	democrats
wrong	hijab	straight	dna	**fox**	atheist	weather	beef	state	politics
self	women	homosexuality	humans	news	**evidence**	pollution	slaughter	money	electoral
death	islam	trans	homo sapiens	terror attacks	faith	global warming	kill	pay	us government
karma	conceals	individual right	darwin's theory	plane	irrational	floods	life	federal	obama
suffering	**death**	gay marriages	life	cia	jesus	ice	diet	state taxes	**policy**
afterlife	tyranny	heterosexual	intelligence	conspiracy theory	supreme being	nuclear waste	meat industry	services	elections
self realization	terrorism	legal	mendel's theory	al qaeda	creationism	sea level	vegan	social security	senate
mortal	muslim bigots	**law**	human dna	bin laden	big bang	climate change	vegetables	income tax	libertarian
self knowledge	sexist	sex	theory of evolution	bush	omnipotent	arctic ice	meat consumption	budget	left wing

(b) LDA

t_1: Spirituality	t_2: Burka/Veil	t_3: Homo-sexuality	t_4: Evolution of Life	t_5: 9/11 Attacks	t_6:Theism/Atheism	t_7: Global warming	t_8: Vegetarianism	t_9: IRS/Taxes	t_{10}: U.S. Politics
life	burka	gay	life	9/11	god	earth	kill	tax	vote
evil	**man**	couples	**god**	laden	religion	planet	meat	pay	electoral
live	women	straight	evolution	bush	jesus	ice	animal	agent	obama
knowledge	immoral	sex	**religion**	terror	**desire**	weather	cow	income	house
purpose	muslim	dating	intelligence	dead	**bang**	warming	egg	revenue	political
values	islam	family	human	**twin**	faith	floods	cooking	us	american
natural	sadistic	**funny**	beings	plane	life	level	**earth**	irs	democrats
existence	terrorism	trans	theory	obl	man	change	diet	american	party
goal	bigot	love	sea	**new**	adam	**clean**	milk	**draft**	bill
self	sexist	children	biology	crash	exists	polar	**fur**	fund	bush
sex	**beautiful**	**law**	earth	tower	**being**	climate	**herbivorous**	funding	senate
spirit	conceals	**ancient**	dna	**york**	omnipotent	**waste**	vegan	state	presidential

(c) SLDA

t_1: Spirituality	t_2: Burka/Veil	t_3: Homo-sexuality	t_4: Evolution of Life	t_5: 9/11 Attacks	t_6:Theism/Atheism	t_7: Global warming	t_8: Vegetarianism	t_9: IRS/Taxes	t_{10}: U.S. Politics
life	burka	gay	life	9/11	god	planet	kill	tax	vote
evil	muslim	couples	theory	laden	religion	ice	meat	pay	electoral
spirit	hijab	dating	evolution	plane	theist	earth	chicken	income	senate
knowledge	women	sex	**homo**	*nonsense*	life	level	egg	*valid*	house
values	*incorrect*	**cannot**	earth	terror	islam	*agree*	beef	state	*correct*
agree	bigot	family	human	bush	**cannot**	change	**fur**	irs	american
existence	sexist	*disagree*	**argument**	*incorrect*	belief	floods	diet	*agree*	democrats
live	terrorism	**don't**	sea	crash	**argument**	**point**	**claim**	**think**	*definitely*
correct	burqa	trans	biology	obl	adam	arctic	milk	us	bush
don't	**man**	children	*prove*	*bogus*	creationism	polar	*disagree*	revenue	elections
goal	*nonsense*	**funny**	dna	cia	**your**	*indeed*	**strength**	budget	obama
purpose	**beautiful**	**law**	darwin	tower	jesus	**clean**	cow	**draft**	*disagree*

Table 1: Top terms of 10 topics discovered by JTE (a), LDA (b), and SLDA (c). **Red (bold)** denotes errors and *blue (italics)* denotes contention/agreement terms.

occurs) to train the Max-Ent model. We set the number of topics, $T = 100$ and the number of CA-expression types, $E = 2$ (contention and agreement) as in discussion/debate forums, there are usually two expression types (This hypothesis is further statistically validated in §4.4.1).

Due to space constraints, we are unable to list the topics discovered by all proposed models. As JTE is the basic model (others build over it) and is also closest to LDA and SLDA, we compare the top terms for 10 topics discovered by JTE, LDA and SLDA in Table 1. The top topical terms by other models are not so different. However, we will evaluate all the proposed models quantitatively later using the task of identifying topics (or "points") of contention in each contentious post, which is one of our objectives. From Table 1, we can still observe that JTE is quite effective at discovering topics. Its topical terms are more specific and contain fewer semantic clustering errors (marked red in bold) than LDA and SLDA. For example, owing to the generative process of JTE, it is able to cluster phrases like "homo sapiens", "darwin's theory," and "theory of evolution" in t_4 (Evolution of Life), which makes the topic more specific.

It is important to note that both LDA and SLDA cannot separate topics and CA-expressions because they find only topics. That is why we need joint modeling of topics and CA-expressions. We can see that the topics of SLDA do contain some CA-expressions (marked blue in italics) because SLDA constrains all words in a sentence to be generated from a single topic [21]. Since CA-expressions can co-occur with topical words in a sentence, they are clustered with topics, which is undesirable. Our proposed models solve this problem based on the joint model formulations.

Next we look at the discovered CA-expressions. We first list some top CA-expressions found by JTE in Table 2 for qualitative inspection. Since CA-expressions found by JTE-R and JTE-P were quite similar to those of JTE among the top 30 terms, they are omitted here. However, all three models are quantitatively evaluated in the next sub-section. From Table 2, we see that JTE can discover and cluster many correct CA-expressions, e.g., "I disagree," "I refute" and "completely disagree" in contention; and "I accept," "I agree," and "you're correct" in agreement. It additionally discovers more distinctive expressions beyond those observed in the training data of Max-Ent. For example, we find phrases like "I don't buy your", "I really doubt", "can you prove", "you fail to", and "you have no clue" being clustered in contention and phrases like "valid point", "rightly said", "I do support", and "very well put" clustered in agreement. These newly discovered phrases are marked *blue* (in italics) in Table 2.

Lastly, we note that CA-expressions of JTE do contain some errors marked **red** (in bold). However, this is a common issue with

I, disagree, **don't**, I don't, **claim**, **you**, oppose, **debate**, I disagree, **argument**, reject, I reject, I refute, **your**, I refuse, doubt, nonsense, *I contest*, dispute, **I think**, completely disagree, don't accept, don't agree, *your claim isn't*, incorrect, *hogwash*, ridiculous, *I would disagree*, false, *I don't buy your*, **really**, *I really doubt*, your nonsense, **true**, *can you prove*, argument fails, *you fail to*, **sense**, your assertions, *bullshit*, *sheer nonsense*, **cannot**, *doesn't make sense*, why do you, *you have no clue*, *how can you say*, *do you even*, absolute nonsense, *contradict yourself*, absolutely not, *you don't understand*, …

(a) JTE Contention expressions, $\Phi^{E}_{Contention}$

agree, **I**, correct, yes, true, **do**, accept, **you**, I agree, right, **indeed**, indeed correct, I accept, **claim**, **your**, **point**, are right, **don't**, valid, *I concede*, *is valid*, **your claim**, *you are right*, **not really**, *would agree*, **might**, *agree completely*, **very**, yes indeed, **mean**, you're correct, **completely**, *valid point*, **argument**, proves, *do accept*, support, **said**, agree with you, *I do support*, *rightly said*, **personally**, absolutely, *completely agree*, well put, *very true*, *well said*, *personally agree*, **doesn't necessarily**, exactly, *very well put*, absolutely correct, **probably**, *kudos*, acknowledge, *point taken*, *partially agree*, agree entirely, ...

(b) JTE Agreement expressions, $\Phi^{E}_{Agreement}$

Table 2: Top terms (comma delimited) of two expression types of the JTE model. **Red (bold)** terms denote possible errors. *Blue (italics)* terms are newly discovered; rest (black) were used in Max-Ent training.

n	JTE				JTE-R				JTE-P			
	C		A		C		A		C		A	
	@50	@100	@50	@100	@50	@100	@50	@100	@50	@100	@50	@100
100	0.60	0.61	0.58	0.62	0.62	0.63	0.60	0.62	0.62	0.64	0.68	0.64
200	0.64	0.63	0.62	0.61	0.66	0.65	0.66	0.65	0.64	0.67	0.70	0.66
300	0.68	0.67	0.64	0.63	0.70	0.71	0.70	0.67	0.68	0.73	0.76	0.68
400	0.72	0.71	0.70	0.65	0.74	0.76	0.76	0.70	0.76	0.77	0.78	0.71
500	0.74	0.75	0.74	0.68	0.78	0.79	0.80	0.71	0.82	0.81	0.80	0.73

Table 3: p@50, 100 for Contention (C), Agreement (A) of various models across different numbers (n) of terms in the labeled sets used for Max-Ent training.

all unsupervised topic models for text and the reason is that the objective function of topic models does not always correlate well with human judgments [11]. In our case, the issue is mainly due to unigram CA-expressions like "I", "your", "do", etc., which by itself do not signify contention or agreement but show up due to higher frequencies in the corpus. There are also phrase errors like "doesn't necessarily", "not really", etc. A plausible approach to deal with this is to discover significant *n*-grams based on multi-way contingency tables and statistical tests along with linguistic clues to pre-process and filter such terms. These issues are worth investigating and we defer them to our future work.

4.2 CA-expression Evaluation

We now quantitatively evaluate the discovered CA-expressions by all three proposed models in three ways. We first evaluate them directly and then evaluate them indirectly through a classification task. Lastly, we examine the sensitivity of the performance with respect to the amount of labeled data. In this case, we do not have an external baseline method as existing joint topic models cannot discover CA-expressions (See §5). However, we will show that the interaction models, JTE-R and JTE-P, are superior to JTE.

4.2.1 Evaluation of CA-expression Rankings

Since CA-expressions (according to top terms in Φ^E) produced by JTE, JTE-R, and JTE-P are rankings, we evaluate them using *precision @ n* (*p@n*), which gives the precision at different rank positions. This measure is commonly used to evaluate a ranking when the number of correct items is unknown, which is our case. For computing *p@n*, we also investigated multi-rater agreement. Three judges independently labeled the top *n* terms as correct or incorrect for Contention and Agreement. Then, we marked a term to be correct if all the judges deemed it so which was then used to compute *p@n*. Multi-rater agreement using Fleiss kappa was greater than 0.80 for all *p@n*, which imply perfect agreement. This is understandable because one can almost certainly make out whether a term expresses contention, agreement or none.

Figure 2 shows the precisions of contention and agreement expressions for the top 50, 100, 150, 200 positions (i.e., *p@* 50, 100, 150, 200) in the two rankings. We observe that both JTE-R and JTE-P are much better than JTE. JTE-P produces the best results. We believe the reason is that JTE-P's expression models being pair specific (θ_p^E) can capture CA-expressions better as contention/agreement expressed by an author is almost always directed to some other authors forming author-pairs.

4.2.2 Contention/Agreement Post Classification

We now use the task of classifying a post as being contentious or agreeing to evaluate the discovered CA-expressions. This classification task is also interesting in its own right. However, we should note that our purpose here is not to find the best classifier to classify posts but to indirectly show that the discovered CA-expressions are of high quality as they help to perform the classification better than the standard word *n*-grams and part-of-speech (POS) *n*-gram features for text classification.

To perform this experiment, we randomly sampled 1000 posts from our database and asked our human judges (3 graduate students well versed in English) to classify each of those posts as *contentious*, *agreeing*, or *other*. Judges were made to work in isolation to prevent bias. We then labeled a post as contentious or agreeing if all judges deemed it so. In this way, 532 posts were classified as contentious, 405 as agreeing. We inspected the rest 63 posts which had disagreements. We found that 18 of them were the first posts of threads. We removed them as the first posts of threads usually start the discussions and do not express agreement or contention. For the remaining 45 of them, 13 posts were partly contentious and partly agreeing (e.g., "*Although I doubt your claim, I agree with your point that…*"), and the rest were mostly statements of views without contention or agreement. Since the number of these posts is small (only 45), we did not use them in classification. That is, we considered two mutually exclusive classes (contentious and agreeing) for post classification. We also conducted a labeling agreement study of our judges using Fleiss multi-rater kappa and obtained $\kappa_{Fleiss} = 0.87$ which shows perfect agreement among the judges according to the scale[7] provided in [25]. This high agreement is not unnatural because by reading a post one can almost certainly make out whether it is overall contentious or agreeing.

For supervised learning, a challenging issue is the choice of features. Word and POS *n*-grams are traditional features. We now compare such features with CA-expressions discovered by the proposed models. We used the top 1000 (contention and agreement) terms from the CA-expression rankings as features. Using classification learning algorithms, we compare a learner trained on all word and POS *n*-grams with those trained on CA-expressions induced by our models. We used SVM, Naïve Bayes (NB), and Logistic Regression (LR). For SVM, we used SVM^{light} (http://svmlight.joachims.org) and for NB and LR, we used the WEKA implementations (http://www.cs.waikato.ac.nz/ml/weka). For SVM, we tried linear, RBF, polynomial and sigmoid kernels, but linear kernel performed the best and hence we only report the results of SVM using linear kernel. Linear kernel has also been shown very effective for text classification by many researchers, e.g., [22]. Table 4 reports the accuracy results. Accuracy is appropriate here as the two classes are not skewed and we are interested in both classes. All results were obtained through 10-fold cross-validation. As the major advantage of CA-expressions arise from dimensionality reduction and feature selection, we also compared with two state-of-the-art feature selection schemes: Information Gain (IG) and Chi-Square test (χ^2). We can observe that SVM performed the best among all learners. The accuracy dramatically increases with CA-expression (Φ^E) features. JTE, JTE-R, and JTE-P progressively improve the accuracy beyond those obtained by traditional *n*-gram features. JTE-P performed the best. Feature selection schemes also improved performance but the proposed models outperform feature selection schemes as well. All accuracy improvements are significant (p<0.001) using a two tailed *t*-test over 10-fold CV. This clearly shows that CA-expressions are of high quality. We also experimented with different numbers of top CA-expressions as features to see how they affect the accuracy results (Figure 3). Here only SVM results are reported as it performed best. We observe in Figure 3 that when the number of CA-expressions reaches about 1000, the classification accuracies start to level-off for all models.

In fact, for JTE and JTE-R, since we know the per post distribution of CA-expression type (θ_d^E), we can also classify a post directly without supervised learning. For each post, if the probability mass associated with type contention is higher than that of agreement, it is classified as a contentious post else an agreeing post. The accuracies using this scheme are: JTE: 74.9%, JTE-R: 75.9%, which are respectable. It is understandable they are lower than supervised methods because supervised learning uses a large number of features. Note that JTE-P cannot be used directly here as its CA-expression types are conditioned over author pairs (θ_p^E) rather than documents (θ_p^E) as in JTE and JTE-R.

4.2.3 Effect of Labeled Term Set Size

Having evaluated CA-expressions, we now examine the sensitivity of model performance with respect to the amount of labeled data. In Table 3, we report the *p@*50, 100 values for all models across contention and agreement on different sizes of labeled term sets used for learning the Max-Ent λ parameters. We see that with more labeled terms, the results improve which is intuitive as more

[7] No agreement (κ<0), slight agreement ($0 < \kappa \leq 0.2$), fair agreement ($0.2 < \kappa \leq 0.4$), moderate agreement ($0.4 < \kappa \leq 0.6$), substantial agreement ($0.6 < \kappa \leq 0.8$), and almost perfect agreement for $0.8 < \kappa \leq 1.0$.

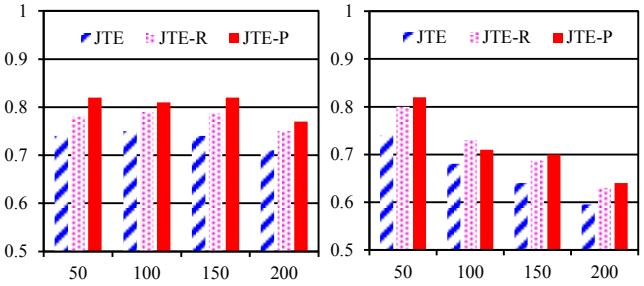


Figure 2: Precision @ top 50, 100, 150, 200 rank positions for Contention (left) and Agreement (right) expressions discovered by various methods.

Features	SVM	NB	LR
W+POS 1-gram	69.37	68.20	68.84
W+POS 1-2 gram	70.33	68.94	69.90
W+POS, 1-3 gram	70.86	69.16	70.44
W+POS, 1-4 gram	70.97	69.26	70.54
W+POS, 1-4 gram + IG	75.67	74.01	75.34
W+POS, 1-4 gram + χ^2	76.21	75.11	76.09
CA-Expr. Φ^E, JTE	80.79	78.55	79.30
CA-Expr. Φ^E, JTE-R	82.18	79.19	80.15
CA-Expr. Φ^E, M-JTE-P	83.88	79.30	81.43

Table 4: Accuracies of post classification: different learners and feature settings. The improvements of our models are significant (p<0.001) over two tailed t-test across 10-fold cross validation.

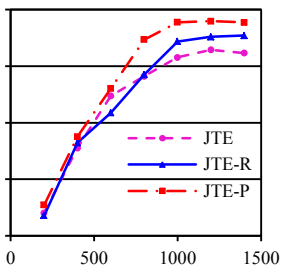


Figure 3: 10-fold CV accuracy of SVM for post classification versus the number of top expressions of Φ^E across all proposed models.

D	Φ^E + Noun/Noun Phrase						JTE						JTE-R						JTE-P					
	J_1			J_2			J_1			J_2			J_1			J_2			J_1			J_2		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
D1	0.60	0.70	0.65	0.55	0.62	0.58	0.75	0.79	0.77	0.68	0.74	0.71	0.78	0.82	0.80	0.74	0.75	0.74	0.79	0.84	0.81	0.74	0.76	0.75
D2	0.61	0.68	0.64	0.57	0.61	0.59	0.74	0.77	0.75	0.67	0.70	0.68	0.76	0.80	0.78	0.70	0.70	0.70	0.77	0.82	0.79	0.70	0.72	0.71
D3	0.62	0.69	0.65	0.54	0.60	0.57	0.73	0.80	0.76	0.65	0.67	0.66	0.77	0.81	0.79	0.71	0.69	0.70	0.77	0.84	0.80	0.72	0.70	0.71
D4	0.63	0.71	0.67	0.53	0.58	0.55	0.72	0.81	0.76	0.63	0.71	0.67	0.74	0.83	0.78	0.67	0.72	0.69	0.73	0.82	0.77	0.68	0.72	0.69
Avg.	0.62	0.70	0.65	0.55	0.60	0.57	0.74	0.79	0.76	0.66	0.71	0.68	0.76	0.82	0.79	0.71	0.72	0.71	0.77	0.83	0.79	0.71	0.73	0.72

Table 5: Evaluation of topics or "points" of contention expressed in posts. For each method, we report the precision (P) and recall (R) for discovering points of contention in posts belonging to a particular domain. The experiments were performed on four domains D1: Politics, D2: Religion, D3: Society, D4: Science. The precision and recall for each domain are the average precision and recall over 125 posts in that domain.

Statistical significance: Differences between Nearest Noun Phrase and JTE for both judges (J_1, J_2) across all domains were significant at 98% confidence level (p<0.02). Differences among JTE, JTE-R and JTE-P for both judges (J_1, J_2) across all domains were significant at 95% confidence level (p<0.05). A two tailed t-test was used for testing significance.

labeled data will result in more accurate Max-Ent estimates. We used 500 labeled terms in all our experiments.

4.3 Discovering Points of Contention

We now turn to the task of automatically discovering points of contention in contentious posts. By "points", we mean the topical terms on which the contention has been expressed. We employ the JTE and JTE-R models in the following manner using estimated θ_d^T. Note that JTE-P cannot be directly used for this task because it has θ^T placed in the author pair plate so its topics are pair specific (θ_p^T) rather than post specific. However, since we know the posterior topic assignments of z_n^d, we can get a posterior estimate of θ_d^T for JTE-P using $\theta_{d,t}^T = \frac{|\{j|z_j^d=t, 1\le j\le N_d\}|}{|\{j|r_{d,j}=\hat{t}, 1\le j\le N_d\}|}$.

Given a contentious post d, we first select the top k topics that are mentioned in d according to its topic distribution, θ_d^T. Let T_d denote the set of these top k topics in d. Then, for each contentious expression $e \in d \cap \varphi_{Contention}^E$, we emit the topical terms of topics in T_d which appear within a word window of v from e in d. More precisely, we emit the set $H = \{w | w \in d \cap \varphi_t^T, t \in T_d, |posi(w) - posi(e)| \le v\}$, where $posi(\cdot)$ returns the position index of the word/phrase in a document d. To compute the intersection $w \in d \cap \varphi_t^T$, we need a threshold. This is so because the Dirichlet distribution has a smoothing effect which assigns some non-zero probability mass to every term in the vocabulary for each topic t. So for computing the intersection, we considered only terms in φ_t^T which have $p(v|t) = \varphi_{t,v}^T > 0.001$ as probability masses lower than 0.001 are more due to the smoothing effect of the Dirichlet distribution than true correlation. In an actual application, the values for k and v can be set according to the user's need. In this experiment, we use k = 3 and v = 5, which are reasonable because a post normally does not talk about many topics (k), and the contention points (topical terms) appear close to the contentious expressions.

For comparison, we also designed a baseline. For each contentious expression $e \in d \cap \varphi_{Contention}^E$, we emit the nouns and noun phrases within the same window v as the points of contention in d. This baseline is reasonable because topical terms are usually nouns and noun phrases and are near contentious expressions. But we should note that this baseline cannot standalone as it has to rely on the expression models Φ^E of JTE-P.

Next, to evaluate the performance of these methods in discovering points of contention, we randomly selected 125 contentious posts from each domain in our dataset and employed the aforementioned methods on the posts to discover the points of contention in each post. Then we asked two human judges (graduate students fluent in English) to manually judge the results produced by each method for each post. We asked them to report the precision (% of terms discovered by a method which are indeed valid points of contention in a post) and recall (% of all valid points of contention which were discovered) for each post. In Table 5, we report the average precision and recall for 125 posts in each domain by the two human judges J_1 and J_2 for different methods. Since this judging task is subjective, the differences in the results from the two judges are not surprising. We observe that across all domains, JTE, JTE-R and JTE-P progressively improve performance over the baseline. Note that it is difficult to compute agreement of two judges using kappa because although the models identify topic terms (which are the same for both judges), the judges also identify additional terms (for recall calculation) which are not found by the models.

4.4 Statistical Experiments

We now compare the proposed models across statistical metrics: perplexity and KL-Divergence.

4.4.1 Perplexity

To measure the ability of JTE, JTE-R and JTE-P to act as good "generative" models, we computed the test-set (see below) perplexity under estimated parameters and also compared with the resulting values of LDA and SLDA models.

Perplexity, widely used in the statistical language modeling community to assess the predictive power of a model, is algebraically equivalent to the inverse of the geometric mean per-word likelihood. A lower perplexity score indicates a better generalization performance. As perplexity monotonically

Iteration	LDA	SLDA	JTE	JTE (E=3)	JTE (E=4)	JTE-R	JTE-P
1000	1795	1745	1475	1489	1497	1467	1442
2000	1684	1568	1381	1397	1423	1343	1326
3000	1575	1474	1318	1331	1348	1273	1257
4000	1561	1421	1248	1266	1278	1223	1208

(a) Perplexity vs. Gibbs Iteration

KL-Div.	LDA	SLDA	JTE	JTE-R	JTE-P
Θ^T	3.4	3.3	9.2	9.9	13.6
Θ^E	-	-	14.1	15.1	17.8
Φ^T	16.8	17.7	20.1	21.8	22.2
Φ^E	-	-	10.9	11.3	11.8

(b) Avg. KL-Div. between models

Table 6: (a) Perplexity comparison of models across Gibbs iterations. The number of topics and CA-expression types were fixed at T = 100, E = 2. All differences are statistically significant (p<0.001) over two tailed t-test across samples from different chains for each group. (b) Average KL-Divergence of topics and CA-expressions, $D_{KL}(\varphi_z||\varphi_{\hat{z}})$ and per document distributions of topics and CA-expressions, $D_{KL}(\theta_d||\theta_{\hat{d}})$. For JTE-P we report the average per pair distributions of topics and CA-expressions $D_{KL}(\theta_p||\theta_{\hat{p}})$. All differences are significant (p<0.01) over two tailed t-test.

decreases with increase in log-likelihood (by definition), it implies that lower perplexity is better since higher log-likelihood on training data means that the model "fits" the data better and a higher log-likelihood on the test set implies that the model can "explain" the unseen data better. Thus, lower perplexity implies that the words are "less surprising" to the model. In our experiments we used 15% of our data (in §4) as our held out test set. As JTE-P requires pair structures, for proper comparison across all models, the corpus was restricted to posts which have at least one quotation. This is reasonable as quoting/replying is an integral part of debates/discussions and we found that about 77% of all posts have quoted/replied-to at least one other post (this count excludes the first posts of threads as they start the discussions and usually have nobody to quote/reply-to). The perplexity (*PPX*) of JTE given the learned model parameters Φ^T, Φ^E and the state of the Markov chain $\mu = \{\vec{z}, \vec{r}, \vec{w}\}$ is given by:

$$PPX = exp\left[-\frac{1}{\overline{W}}\sum_{\tilde{d}=1}^{D_{test}}\frac{1}{S}\sum_{s=1}^{S} logP(\tilde{d}|\mu^s)\right] \quad (8)$$

where $\overline{W}$ denotes the total number of terms in D_{test}. To obtain a better estimate, we average the per-word log-likelihood over S different chains. μ^s denotes the Markov state corresponding to chain s. From the generative process of JTE, we get:

$$logP(\tilde{d}|\mu^s) = \sum_{v=1}^{V}\left(n_{\tilde{d}}^{v}\log\left(\sum_{t=1}^{T}\varphi_{t,v}^{s}\theta_{\tilde{d},t}^{s} + \sum_{e=1}^{E}\varphi_{e,v}^{s}\theta_{\tilde{d},e}^{s}\right)\right) \quad (9)$$

where $n_{\tilde{d}}^{v}$ denotes the number of times term $v \in V$ occurred in $\tilde{d} \in D_{test}$ and $\theta_{\tilde{d},t}^{s}$ and $\theta_{\tilde{d},e}^{s}$ are estimated by querying the model according to the query sampler. In a similar way, the perplexities of the other three models can also be derived.

We compare model perplexities across Gibbs iterations with $T = 100$ and $E = 2$ in Table 6 (a). We note the following observations: i) *Noise Reduction*: The proposed models attain significantly (see caption of Table 6(a)) lower perplexities with fewer iterations than LDA and SLDA showing that the new models fit the debate/discussion forum data better and clustering using the new framework contains less noise. This improvement is attributed to the capabilities of the framework to separate and account for CA-expressions using Φ^E. ii) *The number of CA-expression types*, E: In §2 and 4, we hypothesized that for discussions/debates we mostly have two expression types: contention and agreement. To test this hypothesis, we ran JTE with E > 2 (Table 6(a), columns 5, 6). We find that the test-set perplexity slightly increases (than E = 2) showing performance degradation. It is interesting to note that the number of CA-expression types E impacts perplexity differently than the number of topics T (as the decrease in perplexity usually slows in inverse proportions with increase in the number of topics [4]). We also tried increasing E to 5, 10, etc. However, the performance deteriorated with increase in model perplexity. This result supports our prior hypothesis of E = 2 in debate forums.

4.4.2 KL-Divergence

Another important measure for topic models is topic distinctiveness [24]. Here, we want to assess how distinctive the discovered topics and CA-expressions are. To measure topic and CA-expression distinctiveness, we computed the average topic and CA-expression distribution (φ_z^T and φ_z^E) separations between all pairs of latent topics and CA-expression types. To measure separations, we choose KL-Divergence as our metric as suggested in [24]. Clearly, for more distinctive topic and CA-expression discovery, it is desirable to have higher average KL-Divergence. Table 6(b) shows the comparison results. Again, as JTE-P requires pair structures, for proper comparison, all models were run on the restricted corpus where posts have at least one quotation. We observe that topics discovered by the new models are more distinctive than LDA and SLDA. For both topics and CA-expressions, JTE-R performed better than JTE showing that reply relations are highly beneficial. JTE-P with pair structures performed the best. Table 6(b) also reports the average separations of per document distribution of topics and expressions (θ_d^T and θ_d^E). For models JTE and JTE-R, having higher average KL-Divergence for Θ^T and Θ^E implies that documents are well separated based on the estimated topics and two CA-expression types exhibited. We see that both JTE and JTE-R obtain higher average KL-Divergence for Θ^T than LDA and SLDA. Such separations are particularly useful when topic models are used for performing retrieval [19].

Lastly, we look at the average per pair separation of topics and CA-expressions for JTE-P. Clearly, the KL-Divergence values indicate good separations. This information may be further used to mine contending author-pairs or classify these pairs according to interests, i.e., the posterior on θ_p^E can be used to make predictions on the interaction nature of any two authors. However, these are beyond the scope of this paper. We will study them in future.

5. RELATED WORK

Although limited research has been done on fine-grained mining of contentions in discussion/debate forums, there are several general research areas that are related to our work.

Sentiment analysis: Sentiment analysis mines positive and negative opinions from text [27,34]. Agreement and contention are different concepts. Sentiments are mainly indicated by sentiment words (e.g., *great*, *good*, *bad*, and *poor*), while contention and agreement are indicated by CA-expressions. Sentiment analysis approaches are thus not directly applicable to our tasks and we also need to discover CA-expressions. However, from a modeling perspective, there exist several joint sentiment and topic models which are related to our work and are discussed below. Other related works in sentiment analysis include pro-con classification [3], contradictions [23, 36], attitude [38] and negotiations [39].

Topic models: Topic models such as Latent Dirichlet Allocation (LDA) [4] have been shown effective in mining topics in large text collections. There have been many extensions to correlated [5], supervised [7], multi-grain [43] and sequential [13] topic models. In the context of our JTE-R model, constraining document-topic distributions of a post d to be similar to its quoted posts in q_d is related to topic modeling with network structure [12, 17, 28, 33] where for each pair of documents, a binary link variable is conditioned on their contents. This requires sampling all D^2 links which can be computationally very expensive for large corpora. Chang and Blei [10] improved the efficiency by treating non-links as hidden. In [40], links were assumed to be fixed and topics were conditioned on both the document itself and other linked documents. [29] used a network regularization approach to ensure topics of neighboring documents in a network are similar. This work is based on pLSA [20] and EM rather than LDA and Gibbs sampling. Our approach is simpler. However, all these existing

models are mainly used to find topics for corpus exploration. When applied to discussion/debate forums, they are unable to discover contentious topics and CA-expressions at the same time.

There have also been attempts to jointly model both topics and opinions in sentiment analysis. For example, the ME-LDA model [47] added a sentence plate and used maximum-entropy to separate topics and sentiment terms. [26] added a sentiment variable to LDA and conditioned topics over sentiments. The TSM model [30] encodes positive, negative and neutral sentiment variables and a background word variable to separate topical words from background words. In the ASUM model [21], for each sentence a sentiment is chosen over a multinomial, and a topic is chosen conditioned on the sentiment. However, contention and agreeing expressions are emitted differently. Unlike sentiments and topics which are mostly emitted in the same sentence, contention and agreeing expressions often interleave with users' topical viewpoints and may not be in the same sentence. Most importantly, JTE-R and JTE-P can capture the key characteristics of discussions: topic and author interactions, using reply relations and pair structures. Existing joint models are unable to use them.

Support/oppose classification: There have been works aimed at putting debate authors into support/oppose camps. In [15], conversations are classified into agree, disagree, backchannel and other classes. In [1], the reply network was used to classify discussion participants into camps. In [32], a rule-based approach first classifies replies into contention and agreement classes, and max-cut then classifies authors into opposite camps. None of these works, however, mines CA-expressions or contentious topics.

Stances in online debates: In [40], an unsupervised classification method was proposed to recognize stances in debates. In [42], speaker stances were mined using a SVM classifier. In [8], collective classification techniques were employed. Clearly, our work is different as these existing classification methods do not mine CA-expressions or points of contention in each post.

6. CONCLUSIONS

This paper proposed the task of mining contentions in discussion/debate forums, and presented three new models as a principled way to jointly model and mine topics and linguistic expressions indicating agreements and contentions considering topic and author interactions. Existing models are unable to perform these tasks. Specifically, a joint model (JTE) of topics and CA-expressions was first proposed, which was then improved by encoding two key features of discussions or debates, i.e., topical interactions (using reply-to relations) and author interactions (through pair structures), which yielded the JTE-R and JTE-P models. Experimental results showed that the proposed models outperformed baselines for our tasks: i) discovering topics and CA-expressions; and ii) for each contentious post, discovering the contention points or topics. Statistical experiments of perplexity and KL-Divergence were also conducted. They showed that the proposed models fit the data better and discover more distinctive topics and CA-expressions. In all experiments, the interaction models JTE-R and JTE-P consistently gave better results.

7. ACKNOWLEDGMENT

This work is supported in part by National Science Foundation (NSF) under grant no. IIS-1111092.

8. REFERENCES

[1] Agarwal, R. Rajagopalan, S. Srikant, R. Xu. Y. Mining newsgroups using networks arising from social behavior. *WWW*. 2003.
[2] Andrzejewski, D., Zhu, X., Craven, M. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. *ICML*. 2009.
[3] Awadallah, R., Ramanath, M. and Weikum, G. Language-model-based pro/con classification of political text. *SIGIR*. 2010.
[4] Blei, D., Ng, A., Jordan,M. Latent Dirichlet Allocation. *JMLR*. 2003.
[5] Blei, D. and Lafferty, J. D. Correlated topic models. *NIPS*. 2006.
[6] Blei, D. and Lafferty J. Visualizing topics with multi-word expressions. Tech. Report. 2009. arXiv:0907.1013v1.
[7] Blei, D. and McAuliffe, J. Supervised topic models. *NIPS*. 2007.
[8] Burfoot, C., Bird, S., Baldwin, T. Collective Classification of Congressional Floor-Debate Transcripts. *ACL*. 2011.
[9] Canini, K., Shi, L., Griffiths, T. 2009. Online inference of topics with latent Dirichlet allocation. AISTATS. 2009.
[10] Chang, J. and Blei, D. Relational Topic Models for Document Networks. *AISTATS*. 2009.
[11] Chang, J., Boyd-Graber, J., Wang, C. Gerrish, S. Blei, D. Reading tea leaves: How humans interpret topic models. *NIPS*. 2009.
[12] Daume III, H. Markov random topic fields. *ACL-IJCNLP* 2009.
[13] Du, L., Buntine, W. L. and Jin, H. Sequential Latent Dirichlet Allocation: Discover Underlying Topic Structures within a Document. *ICDM*. 2010.
[14] Easley D. and Kleinberg. J. Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge, 2010.
[15] Galley, M., McKeown, K., Hirschberg, J., Shriberg, E. Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies. *ACL*. 2004.
[16] Gerrish, S. and Blei, D. A Language-based Approach to Measuring Scholarly Impact. *ICML*. 2010.
[17] Griber, A., Rosen-Zvi, M., Weiss, Y. Latent Topic Models for Hypertext. *UAI* 2008.
[18] Griffiths, T. and Steyvers, M. Finding scientific topics. *PNAS*. 2004.
[19] Heinrich, G. Parameter estimation for text analysis. Tech. Rep. 2005.
[20] Hofmann, T. Probabilistic Latent Semantic Analysis. *UAI*. 1999.
[21] Jo, Y. and Oh, A. Aspect and sentiment unification model for online review analysis. *WSDM*. 2011.
[22] Joachims, T. (1998) Text categorization with support vector machines: Learning with many relevant features.. *ECML*, 1998.
[23] Kawahara, D., Inui, K. and Kurohashi, S. Identifying contradictory and contrastive relations between statements to outline web information on a given topic. *COLING*. 2010.
[24] Kawamae, N. Latent interest-topic model. *CIKM*. 2010.
[25] Landis, J. R. and Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174, 1977.
[26] Lin, C. and He, Y. Joint sentiment/topic model for sentiment analysis. *CIKM*. 2009.
[27] Liu, B. Sentiment analysis and opinion mining. *Morgan & Claypool Publishers*, 2012.
[28] Liu, Y., Niculescu-Mizil, A., Gryc, W. Topic-Link LDA: Joint Models of Topic and Author Community. *ICML*. 2009.
[29] Mei, Q., Cai, D., Zhang, D., Zhai, C. Topic modeling with network regularization. *WWW*. 2008.
[30] Mei, Q. Ling, X., Wondra, M., Su, H. and Zhai, C. Topic sentiment mixture: modeling facets and opinions in weblogs. *WWW*. 2007.
[31] Mimno, D. and McCallum, A. Expertise modeling for matching papers with reviewers. *KDD*. 2007.
[32] Murakami A., and Raymond, R. Support or Oppose? Classifying Positions in Online Debates from Reply Activities and Opinion Expressions. *COLING*, 2010.
[33] Nallapati, R., Ahmed, A., Xing, E., Cohen, W. Joint Latent Topic Models for Text and Citations. *KDD*. 2008.
[34] Pang, B. and Lee, L. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*. 2008.
[35] Ramage, D., Hall, D., Nallapati, R. and Manning, C. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. *EMNLP*. 2009.
[36] Ritter, A., Downey, D., Soderland, S. and Etzioni, O. It's a contradiction — No, it's not: A case study using functional relation. EMNLP. 2008.
[37] Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smith, P. The author-topic model for authors and documents. *UAI*. 2004.
[38] Sauper, C. Haghighi, A. and Barzilay, R. Content models with attitude. *ACL*, 2011.
[39] Sokolova, M., and Szpakowicz, S. Language Patterns in the Learning of Strategies from Negotiation Texts. CAI, 2006.
[40] Somasundaran, S., Wiebe, J. Recognizing stances in online debates. *ACL-IJCNLP*. 2009.
[41] Sun, C., Gao, B., Cao, Z., Li, H. HTM: A Topic Model for Hypertexts. *EMNLP*. 2008.
[42] Thomas, M., Pang, B., and Lee, L. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. *EMNLP*. 2006.
[43] Titov, I. and McDonald, R. Modeling online reviews with multi-grain topic models. *WWW*. 2008.
[44] Wallach, H. Topic modeling: Beyond bag of words. *ICML*. 2006.
[45] Wang, X., McCallum, A., Wei, X. Topical n-grams: phrase and topic discovery, with an application to information retrieval. *ICDM*. 2007.
[46] Zhao, X., Jiang, J., He, J., Song, Y., Achanauparp, P., Lim, E., Li, X. Topical Keyphrase Extraction from Twitter ACL-HLT. 2011.
[47] Zhao, X., Jiang, J. Yan, H., Li, X. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. *EMNLP*. 2010.