Discovering User Interactions in Ideological Discussions

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

Abstract

Online discussion forums are a popular platform for people to voice their opinions on any subject matter and to discuss or debate any issue of interest. In forums where users discuss social, political, or religious issues, there are often heated debates among users or participants. Existing research has studied mining of user stances or camps on certain issues, opposing perspectives, and contention points. In this paper, we focus on identifying the nature of interactions among user pairs. The central questions are: How does each pair of users interact with each other? Does the pair of users mostly agree or disagree? What is the lexicon that people often use to express agreement and disagreement? We present a topic model based approach to answer these questions. Since agreement and disagreement expressions are usually multi-word phrases, we propose to employ a ranking method to identify highly relevant phrases prior to topic modeling. After modeling, we use the modeling results to classify the nature of interaction of each user pair. Our evaluation results using real-life discussion/debate posts demonstrate the effectiveness of the proposed techniques.

1 Introduction

Online discussion/debate forums allow people with common interests to freely ask and answer questions, to express their views and opinions on any subject matter, and to discuss issues of common interest. 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 agree or disagree and argue with one another. Such ideological discussions on a myriad of social and political issues have practical implications in the fields of communication and political science as they give social scientists an opportunity to study real-life discussions/debates of almost any issue and analyze participant behaviors in a large scale. In this paper, we present such an application, which aims to perform fine-grained analysis of user-interactions in online discussions.

There have been some related works that focus on discovering the general topics and ideological perspectives in online discussions (Ahmed and Xing, 2010), placing users in support/oppose camps (Agarwal et al., 2003), and classifying user stances (Somasundaran and Wiebe, 2009). However, these works are at a rather coarser level and have not considered more fine-grained characteristics of debates/discussions where users interact with each other by quoting/replying each other to express agreement or disagreement and argue with one another. In this work, we want to mine the following information:

1. The nature of interaction of each pair of users or participants who have engaged in the discussion of certain issues, i.e., whether the two persons mostly agree or disagree with each other in their interactions.
2. What language expressions are often used to express agreement (e.g., "I agree" and "you're right") and disagreement (e.g., "I disagree" and "you speak nonsense").

We note that although agreement and disagreement expressions are distinct from traditional sentiment expressions (words and phrases) such as *good*, *excellent*, *bad*, and *horrible*, agreement and disagreement clearly express a kind of sentiment as well. They are usually emitted during interactive exchanges of arguments in ideological discussions. This idea prompted us to introduce the concept of *AD-sentiment*. We define the polarity of agreement expressions as *positive* and the polarity of disagreement expressions as *negative*. We refer agreement and disagreement expressions as *AD-sentiment expressions*, or *AD-expressions* for short. AD-expressions are crucial for the analysis of interactive discussions and debates just as sentiment expressions are instrumental in sentiment analysis (Liu, 2012). We thus regard this work as an extension to traditional sentiment

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 671–681,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

analysis (Pang and Lee, 2008; Liu, 2012).

In our earlier work (Mukherjee and Liu, 2012a), we proposed three topic models to mine contention points, which also extract AD-expressions. In this paper, we further improve the work by coupling an information retrieval method to rank good candidate phrases with topic modeling in order to discover more accurate AD-expressions. Furthermore, we apply the resulting AD-expressions to the new task of classifying the arguing or interaction nature of each pair of users. Using discovered AD-expressions for classification has an important advantage over traditional classification because they are domain independent. We employ a semi-supervised generative model called JTE-P to jointly model AD-expressions, pair interactions, and discussion topics simultaneously in a single framework. With such complex interactions mined, we can produce many useful summaries of discussions. For example, we can discover the most contentious pairs for each topic and ideological camps of participants, i.e., people who often agree with each other are likely to belong to the same camp. The proposed framework also facilitates tracking users' ideology shifts and the resulting arguing nature.

The proposed methods have been evaluated both qualitatively and quantitatively using a large number of real-life discussion/debate posts from four domains. Experimental results show that the proposed model is highly effective in performing its tasks and outperforms several baselines.

2 Related Work

There are several research areas that are related to our work. We compare with them below.

Sentiment analysis: Sentiment analysis determines positive and negative opinions expressed on entities and aspects (Hu and Liu, 2004). Main tasks include aspect extraction (Hu and Liu, 2004; Popescu and Etzioni, 2005), polarity identification (Hassan and Radev, 2010; Choi and Cardie, 2010) and subjectivity analysis (Wiebe, 2000). As discussed earlier, agreement and disagreement are a special form of sentiments and are different from the sentiment studied in the mainstream research. Traditional sentiment is mainly expressed with sentiment terms (e.g., *great* and *bad*), while agreement and disagreement are inferred by AD-expressions (e.g., *I agree* and *I disagree*), which we also call AD-sentiment expressions. Thus, this work expands the sentiment analysis research.

Topic models: Our work is also related to topic modeling and joint modeling of topics and other information as we jointly model several aspects of discussions/debates.

Topic models like pLSA (Hofmann, 1999) and LDA (Blei et al., 2003) have proved to be very successful in mining topics from large text collections. There have been various extensions to multi-grain (Titov and McDonald, 2008), labeled (Ramage et al., 2009), and sequential (Du et al., 2010) topic models. Yet other approaches extend topic models to produce author specific topics (Rosen-Zvi et al., 2004), author persona (Mimno and McCallum, 2007), social roles (McCallum et al., 2007), etc. However, these models do not model debates and hence are unable to discover AD-expressions and interaction natures of author pairs.

Also related are topic models in sentiment analysis which are often referred to as Aspect and Sentiment models (ASMs). ASMs come in two main flavors: Type-1 ASMs discover aspect (or topic) words sentiment-wise (i.e., discovering positive and negative topic words and sentiments for each topic without separating topic and sentiment terms) (e.g., Lin and He, 2009; Brody and Elhadad, 2010, Jo and Oh, 2011). Type-2 ASMs separately discover both aspects and sentiments (e.g., Mei et al., 2007; Zhao et al., 2010). Recently, domain knowledge induced ASMs have also been proposed (Mukherjee and Liu, 2012b; Chen et al., 2013). The generative process of ASMs is, however, different from our model. Specifically, Type-1 ASMs use asymmetric hyper-parameters for aspects while Type-2 assumes that sentiments and aspects are emitted in the same sentence. However, AD-expressions are emitted differently. They are mostly interleaved with users' topical viewpoints and span different sentences. Further, we capture the key characteristic of discussions by encoding pair-wise user interactions. Existing models do not model pair interactions.

In terms of discussions and comments, Yano et al., (2009) proposed the CommentLDA model which builds on the work of LinkLDA (Erosheva et al., 2004). Mukherjee and Liu (2012d) mined comment expressions. These works, however, don't model pair interactions in debates.

Support/oppose camp classification: Several works have attempted to put debate authors into support/oppose camps. Agrawal et al. (2003) used a graph based method. Murakami and Raymond (2010) used a rule-based method. In (Galley et al., 2004; Hillard et al., 2003), speaker utterances were classified into agreement, disagreement and backchannel classes.

Stances in online debates: Somasundaran and Wiebe (2009), Thomas et al. (2006), Bansal et al. (2008), Burfoot et al. (2011), and Anand et al. (2011) proposed methods to recognize stances in online debates. Some other research directions include subgroup detection (Abu-Jbara et al., 2012), tolerance analysis (Mukherjee et al., 2013), mining opposing perspectives (Lin and Hauptmann, 2006), linguistic accommodation (Mukherjee and Liu, 2012c), and contention point mining (Mukherjee and Liu, 2012a). For this work, we adopt the JTE-P model in (Mukherjee and Liu, 2012a), and make two major advances. We propose a new method to improve the AD-expression mining and a new task of classifying pair interaction nature to determine whether each pair of users who have interacted based on replying relations mostly agree or disagree with each other.

3 Model

We now introduce the JTE-P model with additional details. JTE-P is a semi-supervised generative model motivated by the joint occurrence of expression types (*agreement* and *disagreement*), topics in discussion posts, and user pairwise interactions. Before proceeding, we make the following observation about online discussions.

In a typical debate/discussion post, the user (author) mentions a few topics (using semantically related topical terms) and expresses some viewpoints with one or more AD-expression types (using agreement and disagreement expressions). AD-expressions are directed towards other user(s), which we call *target*(s). In this work, we focus on explicit mentions (i.e., using @name or quoting other authors' posts). In our crawled dataset, 77% of all posts exhibit explicit quoting/reply-to relations excluding the first posts of threads which start the discussions and usually have nobody to quote/reply-to. Such author-target exchanges usually go back and forth between pairs of users populating a thread of discussion. The discussion topics and AD-expressions emitted are thus caused by the author-pairs' topical interests and their nature of interaction (agreeing vs. disagreeing).

In our discussion data obtained from Volconvo.com, we found that a pair of users typically exhibited a dominant arguing nature (agreeing vs. disagreeing) towards each other across various topics or threads. We believe this is because our data consists of topics like elections, theism, terrorism, vegetarianism, etc. which are often heated and attract people with pre-determined, strong, and polarized stances[1].

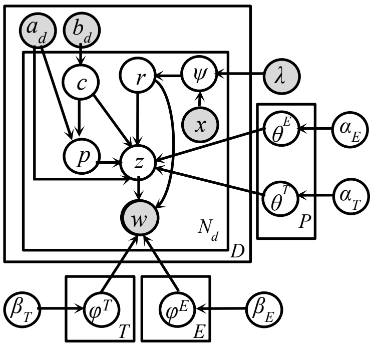

Figure 1: JTE-P Model in plate notation.

Variable/Function	Description
$d; a_d$	A document (post) d ; author a of document, d
$b_d = [b_1 \dots b_n]$	List of *targets* to whom a_d replies/quotes in d.
$p = (a, a')$	Pair of two authors interacting by reply/quote.
θ_p^T; $\theta_p^E(\theta_{p,Ag}^E$, $\theta_{p,DisAg}^E)$	Pair p 's distribution over topics ; expression types (Agreement: $\theta_{p,Ag}^E$, Disagreement: $\theta_{p,DisAg}^E$)
φ_t^T; $\varphi_{e\in\{Ag,DisAg\}}^E$	Topic t 's ; Expression type e 's distribution over vocabulary terms
$T; E$	Total number of topics; expression types
$V; P$	Total number of vocabulary terms; pairs
$w_{d,j}; N_d$	j^{th} term in d; Total # of terms in d
$\psi_{d,j}$	Distribution over topics and AD-expressions
$x_{d,j}$	Associated feature context of the observed term $w_{d,j}$
λ	Learned Max-Ent parameters
$r_{d,j} \in \{\hat{t}, \hat{e}\}$	Binary indicator/switch variable (topic ($\hat{t}$) or AD-expression ($\hat{e}$)) for $w_{d,j}$
$z_{d,j}$	Topic/Expression type of $w_{d,j}$
$\alpha_T; \alpha_E; \beta_T; \beta_E$	Dirichlet priors of θ_p^T; θ_p^E; φ_t^T; φ_e^E
$n_{p,t}^{PT}; n_{p,e}^{PE}$	# of times topic t; expression type e assigned to p
$n_{t,v}^{CT}; n_{e,v}^{CE}$	# of times term v appears in topic t; expression type e

Table 1: List of Notations

This observation motivates the generative process of our model. Referring to the notations in Table 1, we explain the generative process of JTE-P. Given a document (post) d, its author, a_d, and the list of *targets* to whom a_d replies/quotes

[1] These hardened perspectives are supported by theoretical studies in communications like the polarization effect (Sunstein, 2002), and the hostile media effect, a scenario where partisans rigidly hold on to their stances (Hansen and Hyunjung, 2011).

in d, $b_d = [b_1 \dots b_n]$, the document d exhibits shared topics and arguing nature of various pairs, $p = (a_d, c)$, where $c \in b_d$. More precisely, the pair specific topic and AD-expression distributions (θ_p^T ; θ_p^E) "shape" the topics and AD-expressions emitted in d as agreement and disagreement on topical viewpoints are directed towards certain target authors. Each topic (φ_t^T) and AD-expression type (φ_e^E) is characterized by a multinomial distribution over terms (words/phrases). Assume we have $t = 1 \dots T$ topics and $e = 1 \dots E$ expression types in our corpus. Note that in our case of discussion/debate forums, we hypothesize E = 2 as in debates, we mostly find two expression types: *agreement* and *disagreement* (more details in §6.1). Like most generative models for text, a post (document) is viewed as a bag of *n*-grams and each *n*-gram (word/phrase) takes one value from a predefined vocabulary. In this work, we use up to 4-grams, i.e., *n* = 1, 2, 3, 4. Instead of using all *n*-grams, a relevance based ranking method is proposed to select a subset of highly relevant *n*-grams for model building (details in §4). For notational convenience, we use *terms* to denote both *words* (unigrams) and *phrases* (*n*-grams).

JTE-P is a switching graphical model (Ahmed and Xing, 2010; Zhao et al., 2010) performing a switch between AD-expressions and topics. $\psi_{d,j}$ denotes the distribution over topics and AD-expressions with $r_{d,j} \in \{\hat{t}, \hat{e}\}$ denoting the binary indicator/switch variable (topic or AD-expression) for the j^{th} term of d, $w_{d,j}$. To perform the switch we use a maximum entropy (Max-Ent) model. The idea is motivated by the observation that topical and AD-expression terms usually play different roles in a sentence. Topical terms (e.g., "elections" and "income tax") tend to be noun and noun phrases while AD-expression terms ("I refute", "how can you say", and "probably agree") usually contain pronouns, verbs, wh-determiners, and modals. In order to utilize the part-of-speech (POS) tag information, we place the topic/AD-expression distribution $\psi_{d,j}$ (the prior over the indicator variable $r_{d,j}$) in the term plate (see Figure 1) and set it from a Max-Ent model conditioned on the observed feature context $x_{d,j}$ associated with $w_{d,j}$ and the learned Max-Ent parameters, λ (details in §6.1). In this work, we use both lexical and POS features of the previous, current, and next POS tags/lexemes of the term $w_{d,j}$ as the contextual information, i.e., $x_{d,j} = [POS_{w_{d,j-1}}, POS_{w_{d,j}}, POS_{w_{d,j+1}}, w_{d,j-1}, w_{d,j}, w_{d,j+1}]$, which is used to produce the feature functions for Max-Ent. For phrasal terms (*n*-grams), all POS tags and lexemes of $w_{d,j}$ are considered as contextual information for computing feature functions in Max-Ent. We now detail the generative process of JTE-P (plate notation in Figure 1) as follows:

```
1. For each AD-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 pair $p$, draw $\theta_p^E \sim Dir(\alpha_E)$; $\theta_p^T \sim Dir(\alpha_E)$
4. For each forum discussion post $d \in \{1 \dots D\}$:
   i. Given the author $a_d$ and the list of targets $b_d$, for
      each term $w_{d,j}, j \in \{1 \dots N_d\}$:
      a. Draw a target $c \sim Uni(b_d)$
      b. Form pair $p = (a_d, c), c \in b_d$
      c. Set $\psi_{d,j} \leftarrow MaxEnt(x_{d,j}; \lambda)$
      d. Draw $r_{d,j} \sim Bern(\psi_{d,j})$
      e. if ($r_{d,j} = \hat{e}$) // $w_{d,j}$ is an AD-expression term
                 Draw $z_{d,j} \sim Mult(\theta_p^E)$
         else // $r_{d,j} = \hat{t}$, $w_{d,j}$ is a topical term
                 Draw $z_{d,j} \sim Mult(\theta_p^T)$
      f. Emit $w_{d,j} \sim Mult(\varphi_{z_{d,j}}^{r_{d,j}})$
```

Dir, $Mult$, $Bern$, and Uni correspond to the Dirichlet, Multinomial, Bernoulli, and Uniform distributions respectively. To learn JTE-P, we employ approximate posterior inference using Monte Carlo Gibbs sampling. Denoting the random variables $\{w, z, p, r\}$ associated with each term by singular subscripts $\{w_k, z_k, p_k, r_k\}$, $k_{1 \dots K}$, $K = \sum_d N_d$, a single Gibbs sweep consists of performing the following sampling.

$$p(z_k = t, p_k = p, r_k = \hat{t} | \dots) \propto \frac{1}{|b_d|} \frac{exp(\sum_{i=1}^{n} \lambda_i f_i(x_{d,j}, \hat{t}))}{\sum_{y \in \{\hat{e}, \hat{t}\}} exp(\sum_{i=1}^{n} \lambda_i f_i(x_{d,j}, y))} \times \frac{n_{p,t_{\neg k}}^{PT} + \alpha_T}{n_{p,(\cdot)_{\neg k}}^{PT} + T\alpha_T} \frac{n_{t,v_{\neg k}}^{CT} + \beta_T}{n_{t,(\cdot)_{\neg k}}^{CT} + V\beta_T} \tag{1}$$

$$p(z_k = e, p_k = p, r_k = \hat{e} | \dots) \propto \frac{1}{|b_d|} \frac{exp(\sum_{i=1}^{n} \lambda_i f_i(x_{d,j}, \hat{e}))}{\sum_{y \in \{\hat{e}, \hat{t}\}} exp(\sum_{i=1}^{n} \lambda_i f_i(x_{d,j}, y))} \times \frac{n_{p,e_{\neg k}}^{PE} + \alpha_E}{n_{p,(\cdot)_{\neg k}}^{PE} + E\alpha_E} \frac{n_{e,v_{\neg k}}^{CE} + \beta_E}{n_{e,(\cdot)_{\neg k}}^{CE} + V\beta_E} \tag{2}$$

Count variables $n_{t,v}^{CT}$, $n_{e,v}^{CE}$, $n_{p,t}^{PT}$, and $n_{p,e}^{PE}$ are detailed in Table 1. Omission of a latter index denoted by (·) represents the marginalized sum over the latter index. $k = (d, j)$ denotes the j^{th} term of document d and the subscript $\neg k$ denotes the counts excluding the term at (d, j). $\lambda_{1 \dots n}$ are the parameters of the learned Max-Ent model corresponding to the *n* binary feature functions $f_{1 \dots n}$ for Max-Ent. These learned Max-Ent λ parameters in conjunction with the observed feature context, $x_{d,j}$ feed the supervision signal for topic/expression switch parameter, *r* which is updated during inference in equations (1) and (2).

4 Phrase Ranking based on Relevance

We now detail our method of pre-processing n-grams (phrases) based on relevance to select a subset of highly relevant n-grams for model building. This has two advantages: (i). A large number of irrelevant n-grams slow inference. (ii). Filtering irrelevant terms in the vocabulary improves the quality of AD-expressions. Before proceeding, we review some existing approaches. Topics in most topic models like LDA are usually unigram distributions. This offers a great computational advantage compared to more complex models which consider word ordering (Wallach, 2006; Wang et al., 2007). This thread of research models bigrams by encoding them into the generative process. For each word, a topic is sampled first, then its status as a unigram or bigram is sampled, and finally the word is sampled from a topic-specific unigram or bigram distribution. This method, however, is expensive computationally and has a limitation for arbitrary length n-grams. In (Tomokiyo and Hurst, 2003), a language model approach is used for bigram phrase extraction.

Yet another thread of research *post-processes* the discovered topical unigrams to form multi-word phrases using likelihood scores (Blei and Lafferty, 2009). This approach considers adjacent word pairs and identifies n-grams which occur much more often than one would expect by chance alone by computing likelihood ratios. While this is reasonable, a significant n-gram with high likelihood score may *not* necessarily be relevant to the problem domain. For instance, in our case of discovering AD-expressions, the likelihood score[2] of p_1 = "the government of" happens to be more than p_2 = "I completely disagree". Clearly, the former is irrelevant for the task of discovering AD-expressions. The reason for this is that likelihood scores or other statistical test scores rely on the relative counts in the multi-way contingency table to compute significance. Since the relative counts of different fragments of the irrelevant phrase p_1, e.g. "the government", and "government of", happen to appear more than the corresponding counts in the contingency table of p_2, the tests assign a higher score. This is nothing wrong per se because the statistical tests only judge significance of an n-gram, but a significant n-gram may not necessarily be relevant in a given problem domain.

Thus, the existing approaches have some major shortcomings for our task. As our goal is to enhance the expressiveness of our models by considering relevant n-grams preserving the advantages of exchangeable modeling, we employ a *pre-processing* technique to rank n-grams based on relevance and consider certain number of top ranked n-grams based on *coverage* (details follow) in our vocabulary. The idea works as follows.

We first induce a unigram JTE-P whereby we cluster the relevant AD-expression unigrams in φ^E_{Ag} and φ^E_{DisAg}. Our notion of *relevance* of AD-expressions is already encoded into the model using priors set from Max-Ent. Next, we rank the candidate phrases (n-grams) using our probabilistic ranking function. The ranking function is grounded on the following hypothesis: a relevant phrase is one whose unigrams are closely related to (or appear with high probabilities in) the given AD-expression type, e : Agreement (Ag) or disagreement ($DisAg$). Continuing from the previous example, given the expression type $\varphi^E_{e=DisAg}$, p_2 is relevant while p_1 is not as "government" and "disagree" are highly unlikely and likely respectively to be clustered in $\varphi^E_{e=DisAg}$. Thus, we want to rank phrases based on $P(Rel = 1|e, p)$ where e denotes the expression type (Agreement/Disagreement), p denotes a candidate phrase. Following the probabilistic relevance model in (Lafferty and Zhai, 2003), we use a similar technique to that in (Zhao et al., 2011) for deriving our relevance ranking function as follows:

$$P(Rel = 1|e,p) = \frac{P(Rel=1|e,p)}{P(Rel=0|e,p)+P(Rel=1|e,p)} = \frac{1}{1+\frac{P(Rel=0|e,p)}{P(Rel=1|e,p)}} = \frac{1}{1+\frac{P(Rel=0,p|\,e)}{P(Rel=1,p|e)}} = \frac{1}{1+\frac{[P(p|Rel=0,e)\times P(Rel=0|e)]}{[P(p|Rel=1,e)\times P(Rel=1|e)]}} \quad (3)$$

We further define $\varepsilon = \frac{P(Rel=0|e)}{P(Rel=1|e)}$. Without loss of generality, one can say that $P(Rel = 0|e) \gg P(Rel = 1|e)$, because there are many more irrelevant phrases than relevant ones, i.e., $\varepsilon \gg 1$. Thus, taking log, from equation (3), we get,

$$\log P(Rel = 1|e,p) = \log\left(\frac{1}{1+\varepsilon\times\frac{P(p|Rel=0,e)}{P(p|Rel=1,e)}}\right) \approx \log\left(\frac{P(p|Rel=1,e)}{P(p|Rel=0,e)}\times\frac{1}{\varepsilon}\right) = \log\left(\frac{P(p|Rel=1,e)}{P(p|Rel=0,e)}\right) - \log\varepsilon \quad (4)$$

Thus, our ranking function actually computes the relevance score $\log\left(\frac{P(p|Rel=1,e)}{P(p|Rel=0,e)}\right)$. The last term, $\log\varepsilon$ being a constant is ignored because it cancels out while comparing candidate n-grams.

[2] Computed using N-gram statistics package, NSP; http://n-gram.sourceforge.net

We now estimate the relevance score of a phrase $p = (w_1, w_2, \dots, w_n)$. Using the conditional independence assumption of words given the indicator variable Rel and expression type e, we have:

$$\log\left(\frac{P(p|Rel=1,e)}{P(p|Rel=0,e)}\right) = \sum_{i=1}^{n} \log\frac{P(w_i|Rel=1,e)}{P(w_i|Rel=0,e)} \quad (5)$$

Given the expression model φ_e^E previously learned by inducing the unigram JTE-P, it is intuitive to set $P(w_i|Rel = 1, e)$ to the point estimate of the posterior on $\varphi_{e,w_i}^E = \frac{n_{e,w_i}^{EV}+\beta_E}{n_{e,(\cdot)}^{EV}+V\beta_E}$, where n_{e,w_i}^{EV} is the number of times w_i was assigned to AD-expression type e and $n_{e,(\cdot)}^{EV}$ denotes the marginalized sum over the latter index. On the other hand, $P(w_i|Rel = 0, e)$ can be estimated using a Laplace smoothed ($\mu = 1$) background model, i.e., $(w_i|Rel = 0, e) = \frac{n_{w_i}+\mu}{n_V+V\mu}$, where n_{w_i} denotes the number of times w_i appears in the whole corpus and n_V denotes the number of terms in the entire corpus.

Next, we throw light on the issue of choosing the number of top k phrases from the ranked candidate n-grams. Precisely, we want to analyze the coverage of our proposed ranking based on relevance models. By coverage, we mean that having selected top k candidate n-grams based on the proposed relevance ranking, we want to get an estimate of how many relevant terms from a sample of the collection were covered. To compute coverage, we randomly sampled 500 documents from the corpus and listed the candidate n-grams[3] in the collection of sampled 500 documents. For this and subsequent human judgment tasks, we use two judges (graduate students well versed in English). We asked our judges to mark all relevant AD-expressions. Agreement study yielded $\kappa_{\text{Cohen}} = 0.77$ showing substantial agreement according to scale [4] provided in (Landis and Koch, 1977). This is understandable as identifying AD-expressions is a relatively easy task. Finally, a term was considered to be relevant if both judges marked it so. We then computed the coverage to see how many of the relevant terms in the random sample were also present in top k phrases from the ranked candidate n-grams. We summarize the coverage results below in Table 2.

k		3000	4000	5000
JTE-P	Agreement	81.34	84.24	87.01
	Disagreement	84.96	87.86	89.64

Table 2: Coverage (in %) of AD-expressions.

We find that choosing top k = 5000 candidate n-grams based on our proposed ranking, we obtain a coverage of 87% for agreement and 89.64 for disagreement expression types which are reasonably good. Thus, we choose top 5000 candidate n-grams for each expression type and add them to the vocabulary beyond all unigrams.

Like expression types $e_{1\dots E}$, we also ranked candidate phrases for topics $t_{1\dots T}$ using $P(Rel = 1|t, p)$. However, for topics, selecting k based on coverage of each topic is more difficult because we induce 50 topics and it is also much more difficult to manually find relevant topical phrases in the sampled data as a topical phrase may belong to more than one topic. We selected top 2000 ranked candidate phrases for each topic using $P(Rel = 1|t, p)$ as we feel that is sufficient for a topic. Note that phrases for topics are not as crucial as for AD-expressions because topics can more or less be defined by unigrams.

5 Classifying Pair Interaction Nature

We now determine whether two users (also called a user pair) mostly agree or disagree with each other in their exchanges, i.e., their pair interaction or arguing nature. This is a relatively new task. We first summarize the closest related works. In (Galley et al., 2004; Hillard et al., 2003; Thomas et al., 2006, Bansal et al., 2008), conversational speeches (i.e., U.S. Congress meeting transcripts) are classified into for or against an issue using various types of features: *durational* (e.g., time taken by a speaker; speech rate, etc.), *structural* (e.g., no. of speakers per side, no. of votes cast by a speaker on a bill, etc.), and *lexical* (e.g., first word, last word, n-grams, etc.). Burfoot et al., (2011) builds on the work of (Thomas et al., 2006) and proposes collective classification using speaker contextual features (e.g., speaker intentions based on vote labels). However, above works do not discover pair interactions (arguing nature) in debate authors. Online discussion forums are textual rather than conversational (e.g., U.S. Congress meeting transcripts). Thus, the *durational*, *structural*, and *contextual* features used in prior works are not directly applicable.

Instead, the model posterior on θ_p^E for JTE-P

[3] These are terms appearing at least 20 times in the entire collection. We do this for computational reasons as there can be many n-grams and n-grams with very low frequency are less likely to be relevant.

[4] No agreement ($\kappa < 0$), 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 $0.8 < \kappa \leq 1.0$.

can actually give an estimate of the overall interaction nature of a pair, i.e., the probability masses assigned to expression types, $e = Ag$(Agreement) and $e = DisAg$ (Disagreement). As $\theta_p^E \sim Dir(\alpha_E)$, we have $\theta_{p,e=Ag}^E + \theta_{p,e=DisAg}^E = 1$. Hence, if the probability mass assigned to any one of the expression types (agreement, disagreement) > 0.5 then according to the model posterior, that expression type is dominant, i.e., if $\theta_{p,Ag}^E > 0.5$, the pair is agreeing else disagreeing.

However, this approach is not the best. As we will see in the experiment section, supervised classification using labeled training data with discovered AD-expressions as features performs better.

6 Empirical Evaluation

We now evaluate the proposed techniques in the context of the JTE-P model. We first evaluate the discovered AD-expressions by comparing results with and without using the phrase ranking method in Section 4, and then evaluate the classification of interaction nature of pairs.

6.1 Dataset and Experiment Settings

We crawled debate/discussion forum posts from Volconvo.com. The forum is divided into various domains. Each domain consists of multiple threads of discussions. For each post, we extracted the post id, author, 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.

Experiment Data: As it is not interesting to study pairs who only exchanged a few posts, we restrict to pairs with at least 20 post exchanges. This resulted in 1241 authors and 1461 pairs. The reduced dataset consists of 1095586 tokens (after n-gram preprocessing in §4), 40102 posts with an average of 27 posts or interactions per pair. Data from all 4 domains are combined for modeling.

Parameter Settings: For all our experiments, we set the hyper-parameters to the heuristic values $\alpha_T = 50/T$, $\alpha_E = 50/E$, $\beta_T = \beta_E = 0.1$ suggested in (Griffiths and Steyvers, 2004). We set the number of topics, $T = 50$ and the number of AD-expression types, $E = 2$ (agreement and disagreement) as in discussion/debate forums, there are usually two expression types[5]. To learn the Max-Ent parameters λ, we randomly sampled 500 terms from the held-out data (10 threads in our corpus which were excluded from the evaluation of tasks in §6.2, §6.3) appearing at least 10 times and labeled them as topical (361) or AD-expressions (139) and used the corresponding features of each term (in the context of posts where it occurs, §3) to train the Max-Ent model.

6.2 AD-Expression Evaluation

We first list some discovered top AD-expressions in Table 3 for qualitative inspection. From Table 3, we can see that JTE-P can cluster many correct AD-expressions, e.g., “I accept”, “I agree”, “you’re correct”, etc. in agreement and “I disagree”, “don’t accept”, “I refute”, etc. in disagreement. In addition, it also discovers and clusters highly specific and more “distinctive” expressions beyond those used in Max-Ent training, e.g., “valid point”, “I do support”, and “rightly said” in agreement; and phrases like “can you prove”, “I don’t buy your”, and “you fail to” in disagreement. Note that terms in black in Table 3 were used in Max-Ent training. The newly discovered terms are marked *blue* in italics. Clustering errors are in **red** (bold).

For quantitative evaluation, topic models are often compared using perplexity. However, perplexity does not reflect our purpose since we are not trying to evaluate how well the AD-expressions in an unseen discussion data fit our learned models. Instead our focus is to evaluate how well our learned AD-expression types perform in clustering semantic phrases of agreement/disagreement. Since AD-expressions (according to top terms in φ^E) produced by JTE-P are rankings, we choose *precision @ n* (*p@n*) as our metric. *p@n* is commonly used to evaluate a ranking when the total number of correct items is unknown (e.g., Web search results, aspect terms in topic models for sentiment analysis (Zhao et al., 2010), etc.). This situation is similar to our AD-expression rankings, φ^E. Further, as $\varphi^E \sim Dir$, the Dirichlet smoothing effect ensures that every term in the vocabulary has some non-zero mass to agreement or disagreement expression type. Thus, it is the ranking of terms in each AD-expression type that matters (i.e., whether the model is able to rank highly relevant terms at the top).

The above method evaluates the original ranking. Another way of evaluating the AD-expression rankings is to evaluate only those newly discovered terms, i.e., beyond those

[5] Values for $E > 2$ were also tried. However, they did not produce any new dominant expression type. There was also a slight increase in the model perplexity showing that values of $E > 2$ do not fit the debate forum data well.

Disagreement expressions ($\varphi^{E}_{e=Disagreement}$)
I, disagree, I don’t, I disagree, **argument**, reject, **claim**, I reject, I refute, **and**, **your**, I refuse, **won’t**, **the claim**, nonsense, *I contest*, dispute, **I think**, completely disagree, don’t accept, don’t agree, incorrect, **doesn’t**, *hogwash*, *I don’t buy your*, *I really doubt*, your nonsense, **true**, *can you prove*, argument fails, *you fail to*, **your assertions**, *bullshit*, *sheer nonsense*, *doesn’t make sense*, *you have no clue*, *how can you say*, *do you even*, *contradict yourself*, …
Agreement expressions ($\varphi^{E}_{e=Agreement}$)
agree, **I**, correct, yes, true, accept, I agree, **don’t**, indeed correct, **your**, I accept, **point**, **that**, I concede, is valid, **your claim**, **not really**, *would agree*, **might**, *agree completely*, yes indeed, absolutely, you’re correct, *valid point*, **argument**, **the argument**, proves, *do accept*, support, agree with you, *rightly said*, **personally**, well put, *I do support*, *personally agree*, **doesn’t necessarily**, exactly, *very well put*, *kudos*, *point taken*, ...

Table 3: Top terms (comma delimited) of two expression types. **Red** (bold) terms denote possible errors. *Blue* (italics) terms are newly discovered; rest (black) terms have been used in Max-Ent training.

P@n	JTE-P (all terms)						JTE-P (excluding labeled ME terms)					
	Agreement			Disagreement			Agreement			Disagreement		
L	50	100	150	50	100	150	50	100	150	50	100	150
100	0.62	0.63	0.61	0.64	0.62	0.63	0.58	0.56	0.57	0.60	0.59	0.58
200	0.66	0.67	0.65	0.68	0.66	0.67	0.62	0.59	0.60	0.64	0.63	0.62
300	0.70	0.70	0.71	0.70	0.68	0.67	0.66	0.66	0.65	0.66	0.66	0.65
400	0.72	0.72	0.73	0.74	0.71	0.70	0.68	0.67	0.69	0.70	0.68	0.69
500	0.76	0.77	0.75	0.76	0.73	0.74	0.70	0.71	0.70	0.72	0.71	0.70

Table 4: Results using terms based on phrase relevance ranking for P @ n= 50, 100, 150 across 100, 200, …, 500 labeled examples (L) used for Max-Ent (ME) training.

P@n	JTE-P (all terms)						JTE-P (excluding ME terms)					
	Agreement			Disagreement			Agreement			Disagreement		
L	50	100	150	50	100	150	50	100	150	50	100	150
500	0.66	0.69	0.69	0.72	0.70	0.70	0.66	0.65	0.64	0.68	0.66	0.65

Table 5: Results using all tokens (without applying phrase relevance ranking) for *P*@50, 100, 150 and 500 labeled examples were used for Max-Ent (ME) training).

Feature Setting	Agreeing			Disageeing		
	P	R	F_1	P	R	F_1
JTE-P-posterior	0.59	0.61	0.60	0.81	0.70	0.75
W+POS 1-4 grams	0.63	0.66	0.64	0.83	0.82	0.82
W+POS 1-4grams + IG (top 1%)	0.64	0.67	0.65	0.84	0.82	0.83
W+POS 1-4 grams + IG (top 2%)	0.65	0.67	0.66	0.84	0.82	0.83
W+POS 1-4 grams + χ^2 (top 1%)	0.65	0.68	0.66	0.84	0.83	0.83
W+POS 1-4 grams + χ^2(top 2%)	0.64	0.68	0.69	0.84	0.82	0.83
AD-Expressions, Φ^E (top 1000)	0.73	0.74	0.73	0.87	0.87	0.87
AD-Expressions, Φ^E (top 2000)	0.77	0.81	0.78	0.90	0.88	0.89

Table 6: Precision (P), recall (R), and F_1 scores of pair interaction evaluation. Improvements in F_1 using AD-expression features (φ^E) are statistically significant (p<0.01) using paired t-test across 5-fold CV.

labeled terms used in Max-Ent training. For this evaluation, we remove those terms that have been used in Max-Ent (ME) training. We report both results in Table 4. We also studied inter-rater agreement using two judges who independently labeled the top n terms as correct or incorrect. A term was marked correct if both judges deemed it so which was then used to compute $p@n$. Agreement using κ_{Cohen} was greater than 0.78 for all $p@n$ computations implying substantial and good agreements as identifying whether a phrase implies agreement or disagreement or none is an easy task. $P@n$ excluding ME labeled terms (Table 4, second column) are slightly lower than those using all terms but are still decent. This is because $p@n$ excluding ME labeled terms removes many correct AD-expressions used in training.

Further to evaluate the sensitivity of performance on the amount of labeled terms for Max-Ent, we computed $p@n$ across different sizes of labeled terms. Table 4 shows $p@n$ for agreement and disagreement expressions across different sizes of labeled terms (L). We find that more labeled terms improves $p@n$ which is intuitive. We used 500 labeled terms in all our subsequent experiments. The result in Table 4 uses relevance ranking (§4).

We now compare with the performance of the model without using phrase relevance ranking. *P@n* results using all tokens (4356787) are shown in Table 5 (with 500 labeled terms for Max-Ent training). Clearly, *P@n* is lower than in Table 4 (last row; with phrase relevance ranking) because without phrase relevance ranking (Table 5) many irrelevant terms can rank high due to co-occurrences which may not be semantically related. This shows that relevance ranking of phrases is beneficial.

6.3 Pair Interaction Nature

We now evaluate the overall interaction nature of each pair of users. The evaluation of this task requires human judges to read all the posts where the two users forming the pair have interacted. Thus, it is hard to evaluate all 1461 pairs in our dataset. Instead, we randomly sampled 500 pairs (≈ 34% of the population) for evaluation. Two human judges were asked to independently read all the post interactions of 500 pairs and label each pair as overall "disagreeing" or overall "agreeing" or "none". The κ_{Cohen} for this task was 0.81. Pairs were finally labeled as agreeing or disagreeing if both judges deemed them so. This resulted in 320 disagreeing and 152 agreeing pairs. Out of the rest 28 pairs, 10 were marked "none" by both judges while 18 pairs had disagreement in labels. We only focus on the 472 agreeing and disagreeing pairs.

As we have labeled data for 472 pairs, we can treat identifying pair arguing nature as a text classification problem where all interactions between a pair are merged in one document representing the pair along with the label given by judges: agreeing or disagreeing. To compare classification performance, we use two feature sets: (i) standard word + POS 1-4 grams and (ii) AD-expressions from φ^E. We use TF-IDF as our feature value assignment scheme. We also try two well-known feature selection schemes Chi-Squared Test (χ^2) and Information Gain (IG). We use the linear kernel[6] SVM (SVM^{light} system in (Joachims, 1999)) as our text classifier. For feature selection using χ^2 and IG, we use two settings: top 1% and 2% of all features ranked according to the selection metric. Also, for estimated AD-expressions (according to probabilities in φ^E), we experiment with top 1000 and 2000 AD-expressions terms for both agreement and disagreement. We summarize comparison results using 5-fold Cross Validation (CV) with two classes: agreeing and disagreeing in Table 6. JTE-P-posterior represents the method using simply the model posterior on θ_p^E to make the decision (see §5). From Table 6, we can make the following observations.

Predicting agreeing arguing nature is harder than that of disagreeing across all feature settings. Feature selection improves performance. χ^2 and IG perform similarly. AD-expressions, φ^Eyields the best performance showing that the discovered AD-expressions are of high quality and reflect the user pair arguing nature well. Selecting certain top terms in φ^E can also be viewed as a form of feature selection. Although prediction performance using model posterior (JTE-P-posterior) is slightly lower than supervised SVM (Table 6, second row), the F_1 scores are decent. Using the discovered AD-expressions (Table 6, last low) as features renders a statistically significant (see Table 6 caption) improvement over other baseline feature settings. This shows that discovered AD-expressions are useful for downstream applications, e.g., the task of identifying pair interactions.

7 Conclusion

This paper studied the problem of modeling user pair interactions in online discussions with the purpose of discovering the interaction or arguing nature of each author pair and various AD-expressions emitted in debates. A novel technique was also proposed to rank *n*-gram phrases where relevance based ranking was used in conjunction with a semi-supervised generative model. This method enables us to find better AD-expressions. Experiments using real-life online debate data showed the effectiveness of the model. In our future work, we intend to extend the model to account for stances, and issue specific interactions which would pave the way for user profiling and behavioral modeling.

Acknowledgments

We would like thank Sharon Meraz (Department of communication, University of Illinois at Chicago) and Dennis Chong (Department of Political Science, Northwestern University) for several valuable discussions. This work was supported in part by a grant from the National Science Foundation (NSF) under grant no. IIS-1111092.

[6] Other kernels polynomial, RBF, and sigmoid did not perform as well.

References

Abu-Jbara, A., Dasigi, P., Diab, M. and Dragomir Radev. 2012. Subgroup detection in ideological discussions. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2012).

Agrawal, R., Rajagopalan, S., Srikant, R., and Xu. Y. 2003. Mining newsgroups using networks arising from social behavior. In Proceedings of the International Conference on World Wide Web (WWW-2003).

Ahmed, A and Xing, E. 2010. Staying informed: supervised and semi-supervised multi-view topical analysis of ideological perspective. In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP-2010).

Anand, P., Walker, M., Abbott, R., Tree, J., Bowmani, R., and Minor, M. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis.

Bansal, M., Cardie, C., and Lee, L. 2008. The power of negative thinking: Exploiting label disagreement in the min-cut classification framework. In Proceedings of the International Conference on Computational Linguistics (Short Paper).

Blei, D., Ng, A., and Jordan, M. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research.

Blei, D. and Lafferty J. 2009. Visualizing topics with multi-word expressions. Tech. Report. arXiv:0907.1013v1.

Brody, S. and Elhadad, S. 2010. An Unsupervised Aspect-Sentiment Model for Online Reviews. In Proceedings of the Annual Conference of the North American Chapter of the ACL (NAACL-2010).

Burfoot, C., Bird, S., and Baldwin, T. 2011. Collective Classification of Congressional Floor-Debate Transcripts. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2001).

Chang, J., Boyd-Graber, J., Wang, C. Gerrish, S. Blei, D. 2009. Reading tea leaves: How humans interpret topic models. In Proceedings of the Neural Information Processing Systems (NIPS-2009).

Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., Ghosh, R. 2013. Leveraging Multi-Domain Prior Knowledge in Topic Models. In Proceedings of the International Joint Conference in Artificial Intelligence (IJCAI-2013).

Choi, Y. and Cardie, C. 2010. Hierarchical sequential learning for extracting opinions and their attributes (Short Paper). In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2010).

Du, L., Buntine, W. L., and Jin, H. 2010. Sequential Latent Dirichlet Allocation: Discover Underlying Topic Structures within a Document. In Proceedings of the IEEE International Conference on Data Mining (ICDM-2010).

Erosheva, E., Fienberg, S. and Lafferty, J. 2004. Mixed membership models of scientific publications. In Proceedings of the National Academy of Sciences (PNAS-2004).

Galley, M., McKeown, K., Hirschberg, J., and Shriberg, E. 2004. Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2004).

Griffiths, T. and Steyvers, M. 2004. Finding scientific topics. In Proceedings of the National Academy of Sciences (PNAS-2004).

Hansen, G. J., and Hyunjung, K. 2011. Is the media biased against me? A meta-analysis of the hostile media effect research. Communication Research Reports, 28, 169-179.

Hillard, D., Ostendorf, M., and Shriberg, E. 2003. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2003).

Hassan, A. and Radev, D. 2010. Identifying text polarity using random walks. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2010).

Hofmann, T. 1999. Probabilistic latent semantic analysis. In Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI-1999).

Hu, M. and Liu, B. 2004. Mining and summarizing customer reviews. In Proceedings of the SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004).

Jo, Y. and Oh, A. 2011. Aspect and sentiment unification model for online review analysis. In Proceedings of the International Conference on Web Search and Data Mining (WSDM-2011).

Joachims, T. Making large-Scale SVM Learning Practical. 1999. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

Lafferty, J. and Zhai, C. 2003. Probabilistic relevance models based on document and query generation. Language Modeling and Information Retrieval.

Landis, J. R. and Koch, G. G. 1977. The measurement of observer agreement for categorical data. Biometrics.

Lin, C. and He, Y. 2009. Joint sentiment/topic model for sentiment analysis. In Proceedings of the

International Conference on Knowledge Management (CIKM-2009).

Lin, W. H., and Hauptmann, A. 2006. Are these documents written from different perspectives?: a test of different perspectives based on statistical distribution divergence. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2006).

Liu, B. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publisher, USA.

McCallum, A., Wang, X., and Corrada-Emmanuel, A. 2007. Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email. Journal of Artificial Intelligence Research.

Mei, Q., Ling, X., Wondra, M., Su, H., and Zhai, C. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In Proceedings of the International Conference on World Wide Web (WWW-2007).

Mimno, D. and McCallum, A. 2007. Expertise modeling for matching papers with reviewers. In Proceedings of the SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2007).

Mukherjee, A., Venkataraman, V., Liu, B., Meraz, S. 2013. Public Dialogue: Analysis of Tolerance in Online Discussions. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2013).

Mukherjee, A. and Liu, B. 2012a. Mining Contentions from Discussions and Debates. Proceedings of SIGKDD Conference on Knowledge Discovery and Data Mining (KDD-2012).

Mukherjee, A. and Liu, B. 2012b. Aspect Extraction through Semi-Supervised Modeling. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2012).

Mukherjee, A. and Liu, B. 2012c. Analysis of Linguistic Style Accommodation in Online Debates. In Proceedings of the International Conference on Computational Linguistics (COLING-2012).

Mukherjee, A. and Liu, B. 2012d. Modeling Review Comments. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2012).

Murakami A. and Raymond, R. 2010. Support or Oppose? Classifying Positions in Online Debates from Reply Activities and Opinion Expressions. In Proceedings of the International Conference on Computational Linguistics (Coling-2010).

Pang, B. and Lee, L. 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval.

Popescu, A. and Etzioni, O. 2005. Extracting product features and opinions from reviews. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2005).

Ramage, D., Hall, D., Nallapati, R, Manning, C. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2009).

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smith, P. 2004. The author-topic model for authors and documents. In Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI-2004).

Sunstein, C. R. 2002. The law of group polarization. *Journal of political philosophy*.

Somasundaran, S. and Wiebe, J. 2009. Recognizing stances in online debates. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP-2009).

Titov, I. and R. McDonald. 2008. Modeling online reviews with multi-grain topic models. In Proceedings of the International Conference on World Wide Web (WWW-2008).

Thomas, M., Pang, B., and Lee, L. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2006).

Tomokiyo, T., and Hurst, M. 2003. A language model approach to keyphrase extraction. In Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18.

Wallach, H. 2006. Topic modeling: Beyond bag of words. In Proceedings of the International Conference on Machine Learning (ICML-2006).

Wang, X., McCallum, A., Wei, X. 2007. Topical N-grams: Phrase and topic discovery, with an application to information retrieval. In Proceedings of the IEEE International Conference on Data Mining (ICDM-2007).

Wiebe, J. 2000. Learning subjective adjectives from corpora. In Proc. of National Conference on AI (AAAI-2000).

Yano, T., Cohen, W. and Smith, N. 2009. Predicting response to political blog posts with topic models. In Proceedings of the N. American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2009).

Zhao, X., J. Jiang, J. He, Y. Song, P. Achananuparp, E.P. LiM, and X. Li. 2011. Topical keyphrase extraction from twitter. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2011).

Zhao, X., Jiang, J., Yan, H., and Li, X. 2010. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2010).