Mining Contrastive Opinions on Political Texts using Cross-Perspective Topic Model

Yi Fang†, Luo Si†, Naveen Somasundaram†, Zhengtao Yu‡
† Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA
‡ Kunming University of Science and Technology, Kunming, China
†{fangy, lsi, nsomasun}@cs.purdue.edu; ‡ztyu@bit.edu.cn

ABSTRACT

This paper presents a novel opinion mining research problem, which is called Contrastive Opinion Modeling (COM). Given any query topic and a set of text collections from multiple perspectives, the task of COM is to present the opinions of the individual perspectives on the topic, and furthermore to quantify their difference. This general problem subsumes many interesting applications, including opinion summarization and forecasting, government intelligence and cross-cultural studies. We propose a novel unsupervised topic model for contrastive opinion modeling. It simulates the generative process of how opinion words occur in the documents of different collections. The ad hoc opinion search process can be efficiently accomplished based on the learned parameters in the model. The difference of perspectives can be quantified in a principled way by the Jensen-Shannon divergence among the individual topic-opinion distributions. An extensive set of experiments have been conducted to evaluate the proposed model on two datasets in the political domain: 1) statement records of U.S. senators; 2) world news reports from three representative media in U.S., China and India, respectively. The experimental results with both qualitative and quantitative analysis have shown the effectiveness of the proposed model.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*; G.3 [**Probability and Statistics**]: Probabilistic algorithms; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*

General Terms

Algorithms, Experimentation

Keywords

contrastive opinions, opinion mining, topic modeling, opinion retrieval

1. INTRODUCTION

Opinion mining is concerned with extracting and analyzing judgments on various aspects of given items from a set of text documents. It is an important task in information retrieval and data mining as it aims at finding subjective information, which may be more relevant to users than factual information in many applications. While there has been much research in opinion mining, most of them focus on analyzing opinions at the word-level, sentence-level or document-level. This paper studies a novel opinion mining problem which examines opinions at the collection-level with each text collection coming from a different perspective. We refer to the task as Contrastive Opinion Modeling (COM): given any query topic and a set of text collections from multiple perspectives, the task is to demonstrate the difference among the perspectives' opinions on the topic. Specifically, COM discovers the common topics across all the perspectives. For each discovered topic or any ad hoc query topic, the task involves: 1) presenting the opinions from each perspective; 2) quantifying their difference. COM models opinions against a whole collection/perspective which potentially consists of a large number of documents. Therefore, it can answer a wide range of opinion analysis requests about the perspective.

There exists much work on opinion retrieval such as the subtask in TREC Blog track [23]. In the TREC task, a set of opinionated documents are returned and users often have to go through the documents to look for the opinions expressed by the perspective of interest. In COM, opinions (in the form of opinion words) are directly returned in response to the user query. For example, it can address a user request like "what are the respective opinions of U.S., China and India (e.g., from news agencies) on *Dalai Lama* and how much difference among them?". To answer this request, the opinion words are returned like "nonviolent" for U.S., "rebellious" for China and "Holy" for India, and a diversity score is also presented so that users can clearly know the degree of discordance among the perspectives. Thus, COM can provide more direct opinion search than the existing work.

Furthermore, a lot of current opinion mining work focuses on mining review data and solving classification problems. As we go beyond product reviews, only knowing sentiment orientations such as positive, negative and neutral is not enough in many cases. This is especially true in the domain of politics where the wording is often sensitive. For example, with respect to healthcare reform in U.S., a Republican might often say "we want responsible healthcare reform based on private insurance"[1], while a Democrat might of-

---

[1] http://www.gop.gov/solutions/healthcare

ten say "we want universal healthcare reform with a public government-run health insurance agency"[2]. Both statements can be viewed as positive on healthcare reform in general, but the opinion words "responsible" and "private" vs "universal" and "public" reflect their huge difference on the issue. Therefore, in COM, the opinions of interest are represented by opinion words which are directly returned to users.

To tackle the task of contrastive opinion modeling, we propose a novel topic model, called Cross-Perspective Topic (CPT) model. The model simulates the generative process of how opinion words appear in the documents. It not only discovers topics but also models the corresponding opinions across multiple perspectives. In CPT, the generative process of opinion words are separated from the generative process of topic words. As a result, besides the word distribution of each topic, we can obtain the opinion distribution for each topic as well. These distributions manifest the associations between topics and opinions, and enable us to accomplish a variety of opining mining tasks including COM. Our contributions in this paper can be summarized as follows:

1. We define and study a novel opinion mining task: contrastive opinion modeling, which aims to directly find the opinions of multiple perspectives with respect to a given topic and quantify their difference on the topic.

2. We propose a fully unsupervised topic model requiring no labeled data for COM. The proposed model is estimated by the Gibbs Sampling algorithm. The opinions on any ad hoc query can be efficiently determined based on the learned parameters.

3. Based on the proposed model, we define a diversity metric of multiple perspectives on a given topic by the Jensen-Shannon divergence among the individual topic-opinion distributions.

4. We conduct an extensive set of experiments with both qualitative and quantitative evaluations on two datasets in the political domain: 1) statement records of U.S. senators; 2) world news reports from three representative media in U.S., China and India, respectively.

2. RELATED WORK

Opinion mining and sentiment analysis have been extensively studied in the recent years. For a general survey, please refer to [24]. The early work focused on identifying the polarity of opinions at the word level [9], at the sentence level [13] and at the document level [25]. These methods do not consider the dependence of opinions on topics.

In [11], topicality and polarity are first combined together to form the notion of opinion retrieval, i.e., to find opinionated documents about a given topic. One early ranking formula is introduced in [6] as the cross entropy of topics and sentiments under a generative model. In 2006, the Text REtrieval Conference (TREC) introduced a Blog Track with a major task of opinion retrieval [23]. An opinion retrieval system is required to locate blog documents expressing opinions. The opinion retrieval task has been approached as a two-stage task: first, retrieving topically relevant documents, and then reranking the documents by the

opinion scores. One popular method to identify opinionated content is by matching the documents with a sentiment word dictionary and calculating term frequency [38]. There are also some interesting work on modeling the topic and sentiment of documents in a unified way [37]. The task of opinion retrieval here is essentially a document retrieval process, without opinions directly returned in response to a search request. Similar to the TREC effort, NTCIR launched the Opinion Analysis Task [29] in 2007 with multilingual testbeds in Chinese, Japanese, and English. One subtask involves detection of opinionated sentences and opinion fragments within opinionated sentences, which is closer to our task while we directly target on opinion words.

Another body of related research is around feature based opinion mining which identifies opinions about the features or attributes of a product instead of giving an overall evaluation. The early representative work is the association rule mining based method [10], and template extraction based method [28]. However, these product opinion features are highly dependent on the training data sets, and thus the methods are not flexible to deal with ad hoc queries and topics. The same problem is shared with [35]. Our work finds the underlying topics (equivalent to features) automatically through topic models and can be applied to ad hoc queries.

Latent Dirichlet Allocation (LDA) [2] is one of the earliest topic models and many variants of LDA have been proposed. Among them, the Correspondence Latent Dirichlet Allocation (corrLDA) model [1] resembles our model in spirit. However, corrLDA models the joint distribution of images (with Gaussian distribution) and their annotations (with multinomial distribution), while our model targets on the generative process of opinions (with multinomial distributions over both opinions and topic words). Furthermore, our model differentiates the perspectives of documents. Several topic models have been proposed for opinion mining. Topic-Sentiment Model [22] calculates sentiment coverage of documents by jointly modeling the mixture of topics and sentiment predictions. Similarly, the Joint Sentiment Topic model [16] is proposed and can directly predict the sentiment orientation at the document level. Considering the hierarchy structure between objects and their associated aspects, the Multi-Grain Latent Dirichlet Allocation model [32] was proposed to find ratable aspects from global topics. They later proposed Multi-Aspect Sentiment model [31] which summarizes sentiment texts by aggregating on each ratable aspects. Recently, the Aspect and Sentiment Unification Model [12] is proposed to model sentiments toward different aspects of an entity. The major difference of all the above work from ours is that the existing work does not retrieve opinion words on ad hoc queries. In addition, they do not model contrastive opinions and are not able to quantify the difference between perspectives.

On the other hand, current opinion mining work mostly focuses on mining product review data [5], because of the wide availability of review data and their relatively obvious sentiment orientations such as good, bad and so on. In this paper, we move beyond the review data and target on the political domain where merely identifying the opinion polarity is not sufficient. In the recent years, political data are increasingly available to the public from a wide range of sources such as political blogs, news media, user comments,

---

[2]http://en.wikipedia.org/wiki/Public_health_insurance_option

and the Open Government Data Initiative[3]. The emerging political data open new opportunities as well as unique challenges for opinion mining and sentiment analysis, which results in an increased interest in the area. For example, in [30] and [4], they study the language and ideology issues on congressional speeches by investigating the contributions of words on ideology. In [34], topic models are proposed to model the discussions in online political blogs and predicts responses to the blog posts. In [3], the opinion scoring models are constructed to extract statements which best express opinionists' standpoints on certain topics. In [18, 19], statistical models are proposed to identify the political perspective of a document or a collection. All the above works do not either generalize to ad hoc query topics nor model contrastive opinions.

There are several studies conducted on comparing texts or opinions. In [20], a system is presented for analyzing and comparing consumer opinions of competing products. In [15], a weakly-supervised bootstrapping method is proposed to identify comparative questions and entities. A probabilistic model for comparing text collections was previously introduced in [36] for a problem called comparative text mining. Given news articles from different sources (about the same event), the model can extract what is common to all the sources and what is unique to one specific source. The model is extended in [26] to detect cultural differences from people's experiences in various countries. In [14] and [27], they tackle the problem of contrastive summarization, which jointly generates summaries for two entities in order to highlight their differences. However, the above works do not quantify the differences, which makes the differences not measurable or comparable among multiple requests. Moreover, they cannot deal with ad hoc queries.
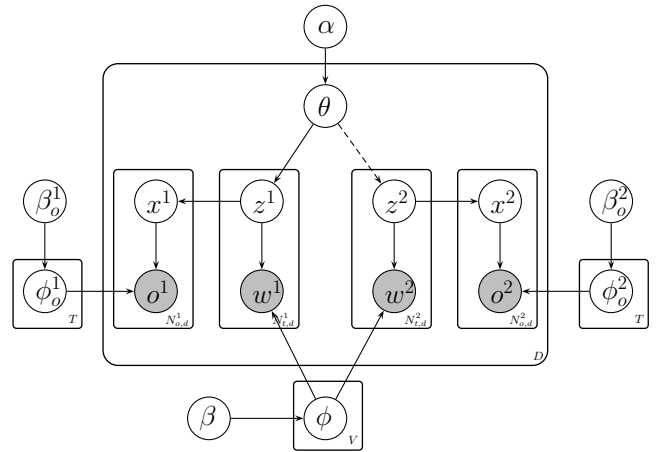
3. CROSS-PERSPECTIVE TOPIC MODEL

Latent Dirichlet Allocation (LDA) [2] is one of the most popular topic models based upon the assumption that documents are mixture of topics, where a topic is a probability distribution over words. LDA is a powerful tool for topic modeling, but it is not well fitted for opinion modeling. A topic in LDA not only contains the words that describe the topic, but also the words that express the opinions about the topic. In other words, LDA does not differentiate opinion words from topic words, which makes both opinions and topics obscure for opinion mining. The problem is even exacerbated when opinions come from multiple different perspectives. In this case, a standard LDA will have severe limitations because it does not directly model the opinions and thus different opinions could be mixed together in the same topics.

In this section, we introduce the Cross-Perspective Topic (CPT) model for contrastive opinion modeling. This model directly depicts how opinions are generated in the documents of different perspectives. In CPT, the opinion generation process is separated from the topic term generation process. We assume that topics are expressed through noun words in the documents, and opinions are conveyed through adjective, verb and adverb words. Section 4.2 gives a detailed description of how to extract the opinion words and topic words from the documents.

The imaginary generative process of the opinions in a

Figure 1: The plate notation of the Cross-Perspective Topic model. The shaded nodes are observed variables.

document is: a person first chooses a topic based on the document, and then she selects a topic word based on the topic. After choosing all the topic words in the document, she selects a topic to express the opinions. The choice of the topic for a opinion word is based on the actual frequency of the topic occurring in the document. Under this topic, she then selects a opinion word based on her perspective. In this model, it is assumed that topics are shared among all the documents, regardless the perspective of the document. Therefore, the topic words are drawn from the shared topic-word distribution. On the other hand, the opinions from different perspective could be different. Thus, the opinion words are drawn from the topic-opinion distribution conditioned on the perspective. Specifically, the topic word w is modeled by a shared LDA across perspectives. The opinion word o is drawn conditioned on the topic x which is uniformly sampled from the topics learned from the topic words in document d. For simplicity of presentation, we only consider two perspectives, but the model can be straightforwardly generalized to more perspectives. The index of the perspective is denoted by the superscript of the variable instance (e.g., w^1 is a topic word in perspective/collection C^1). The generative process in CPT can be described as follows.

1. Draw a perspective-independent multinomial topic word distribution ϕ from Dirichlet(β) for each topic z
2. Draw a perspective-specific multinomial opinion word distribution ϕ_o^i from Dirichlet(β_o^i) for each topic z^i for the perspective C^i
3. For each document d, choose a topic mixture θ from Dirichlet(α)
4. For each topic word w in d
   (a) Draw a topic z from Multinomial(θ)
   (b) Draw a word w from Multinomial(ϕ) conditional on z
5. For each opinion word o in $d \in C^i$,
   (a) draw a topic x^i from Uniform($z_{w_1}, z_{w_2}, ..., z_{w_{N_{t,d}}}$)
   (b) draw a opinion word o^i from Multinomial(ϕ_o^i) conditional on x^i

The graphical model corresponding to this process is shown in Figure 1 with the notations summarized in Table 1. The dashed line from θ to z_2 means that a document can only

come from a single perspective/collection (either C^1 or C^2 in this figure).

3.1 Parameter Estimation

The CPT model has four parameters to estimate: i.e., the document-topic distribution θ, the topic-word distribution ϕ, and the topic-opinion distributions ϕ^1 and ϕ^2. Several methods have been developed for estimating the parameters in LDA, such as Gibbs sampling [8] and variational EM [2]. We employ Gibbs sampling in this paper, because it is comparable in speed to other estimation methods and it approximates a global maximum (whereas EM algorithms may only converge to a local maximum).

Gibbs sampling is a type of Markov chain Monte Carlo algorithm. In a Gibbs sampler, one iteratively samples new assignments of hidden variables by drawing from the distributions conditioned on the previous state of the model. In the Gibbs Sampling procedure of CPT, additional Markov chains are introduced for simulating the opinion generation. We derive the Gibbs sampling equations for our model as follows. The major notations used in the following equations are explained in Table 1.

- Sampling equation of the topic variable z for each topic word w_i:

$$p(z_i = k | w_i = v, \boldsymbol{z}_{-i}, \boldsymbol{w}_{-i}, \alpha, \beta)$$
$$\propto \frac{n_{kd,-i} + \alpha}{\sum_{k=1}^{K} n_{kd,-i} + K\alpha} \times \frac{n_{vk} + \beta}{\sum_{v=1}^{V} n_{vk} + V\beta}$$

- Sampling equation of the opinion topic variable x^1 in the perspective C^1 (the similar equation can be derived for C^2):

$$p(x_i^1 = s | o_i = r, \boldsymbol{x}_{-i}^1, \boldsymbol{o}_{-i}, \beta, \beta_o)$$
$$\propto \frac{n_{rs,-i} + \beta_o^1}{\sum_{r=1}^{T} n_{rs,-i} + T\beta_o^1} \times \frac{n_{sd}}{N_{t,d}}$$

After a set of sampling processes based on the posterior distributions calculated with the above equations, we can estimate the four parameters for any single sample using the following equations:

$$\theta_{kd} = \frac{n_{kd,-i} + \alpha}{\sum_{k=1}^{K} n_{kd,-i} + K\alpha}, \qquad \phi_{vk} = \frac{n_{vk} + \beta}{\sum_{v=1}^{V} n_{vk} + V\beta}$$
$$\phi_{o,rs}^1 = \frac{n_{o,rs} + \beta_o^1}{\sum_{r=1}^{T} n_{o,rs} + T\beta_o^1}, \quad \phi_{o,rs}^2 = \frac{n_{o,rs} + \beta_o^2}{\sum_{r=1}^{T} n_{o,rs} + T\beta_o^2}$$

3.2 Inference for Contrastive Opinion Modeling

The CPT model estimates soft associations between latent topics and observed opinions across different perspectives. These associations are the basis for a number of operations relevant to opinion mining. In this subsection, we present methods to tackle the contrastive opinion modeling problem by utilizing the estimated parameters in the model.

As shown in Introduction, the ad hoc opinion search task in COM is to present the most relevant opinion words to users with respect to a given query topic (and a particular

Table 1: Notations in the Cross-Perspective Topic model

$d, v, r,$ k, s	Instance of a variable: d for document, v for topic word, r for opinion, k for topic of topic word, s for topic of opinion word
D, K	Number of documents and topics
V, T	Size of topic word vocabulary and opinion word vocabulary
$\boldsymbol{w}_{-i}, \boldsymbol{z}_{-i},$ $\boldsymbol{o}_{-i}$	The vector values of $\boldsymbol{w}, \boldsymbol{z}_i$ and $\boldsymbol{o}_i$ on all the other dimensions except i
$N_{t,d}$	Number of topic words in document d
$N_{o,d}$	Number of opinion words in document d
$n_{kd,-i}$	Number of times topic k has occurred in document d, except the current instance
$n_{vk,-i}$	Number of times word v is assigned to topic k, without counting the current instance
$n_{rs,-i}$	Number of times opinion r is assigned to topic s, without counting the current instance
n_{sd}	Number of times topic s occurs in document d
θ	$D \times K$ matrix for document-topic distribution
ϕ	$K \times V$ matrix for topic-word distribution
ϕ_o^1, ϕ_o^2	$K \times T$ matrices for topic-opinion distribution

perspective C^i). Based on the estimated CPT model, this can be done by calculating a predictive likelihood $p(o = r|q, C^i)$ that the opinion word r could be generated by the query q as follows:

$$p(o = r|q, C^i) = \sum_{k=1}^{K} p(o = r|z = k)p(z = k|q)$$
$$= \sum_{k=1}^{K} p(o = r|z = k) \frac{p(q|z = k)p(z = k)}{p(q)}$$
$$\propto \sum_{k=1}^{K} \phi_{o,rk}^i \phi_{qk} n_k \qquad (1)$$

In the above derivation, the Bayes' rule is used, $p(q)$ is constant for the same query, and the unconditional topic probability $p(z = k) \propto n_k$, where n_k is the perspective-wide total number of words associated with topic k. $\phi_{o,rk}^i$ and ϕ_{qk} are topic-opinion distribution and topic-word distribution, respectively, which are the parameters estimated in the CPT training process. The opinion words are then ranked according to the descending order of $p(o = r|q, C^i)$ in response to query q. Intuitively, Eqn. (1) is a weighted inner product between two vectors that penalizes weak topics. Moreover, it is worth noticing that $\phi_{o,rk}^i$, ϕ_{qk} and n_k are all computed offline, which makes the opinion search process very efficient.

As discussed in Section 1, another important aspect of contrastive opinion modeling is to quantify the difference between different perspectives' stance on the topic. In fact, $p(o = r|q, C^1)$ provides the basis for accomplishing the task, because $p(o = r|q, C^1)$ is essentially the probability that perspective C^1 uses the word r to express her opinions on the issue q. The perspectives with similar opinions will have similar $p(o|q)$ and the perspectives with contrary opinions will have very different $p(o|q)$. Therefore, we can compare the distributions of different perspectives, i.e., $p(o|q, C^1)$ vs $p(o|q, C^2)$, to find out their difference.

A natural and well studied "distance" between distribu-

tions is the Kullback-Leibler (KL) divergence. However, the KL-divergence suffers from two drawbacks: 1) it is not symmetric in its arguments and 2) it does not naturally generalize to measuring the divergence among more than two distributions. We instead employ the related Jensen-Shannon divergence [17]. Formally, we define a diversity metric between multiple perspectives on a topic by the Jensen-Shannon divergence. Given a set of query-opinion distributions $\{p(o^1|q), ..., p(o^m|q)\}$ from m perspectives, let $\bar{p}$ be the average (centroid) of these distributions. The Jensen-Shannon divergence JS among these distributions is then defined as the average of the KL-divergences of each distribution to this average distribution as follows:

$$JS(C^1, ..., C^m) = \frac{1}{m} \sum_{j=1}^{m} KL\Big(p(o^j|q)\|\bar{p}\Big)$$

where

$$KL\Big(p(o^j|q)\|\bar{p}\Big) = \sum_{o=1}^{T} p(o^j|q) \log \frac{p(o^j|q)}{\bar{p}}$$

$$\bar{p} = \frac{1}{m} \sum_{j=1}^{m} p(o^j|q)$$

4. EXPERIMENTAL SETUP

4.1 Data Collections

We create two datasets for the evaluation of the proposed model. The first dataset contains the statement records of U.S. senators crawled from the Project Vote Smart[4] website. These statement records present the political stances of senators. The second dataset includes the international headline news published in New York Times[5], Xinhua News[6] and The Hindu[7] during the period of January 2009 - December 2010. The three news agencies are the influential media in US, China, and India, respectively, and usually express representative opinions for these three countries. These world news are all in English and are reported around the same period. The topics covered are thus expected to be largely overlapped. Both datasets were automatically tokenized and sentence split. Table 2 gives detailed statistics of the collections. Before applying the topic models we removed punctuation and also removed stop words using the standard list of stop words[8] [9].

The CPT model has four Dirichlet hyper parameters α, β, β^1 and β^2. Previous research found that these hype parameters only affect the convergence of Gibbs sampling but not much the output results [8]. We fix them to $\alpha = 50/K$ and $\beta = \beta^1 = \beta^2 = 0.02$ according to [8] for all the experiments.

4.2 Opinion Word Extraction

We treat all the nouns in the documents as the topic words. For the opinion words, we use the adjectives, verbs

---

[4]http://www.votesmart.org
[5]http://www.nytimes.com/pages/world
[6]http://www.xinhuanet.com/english2010/world
[7]http://www.thehindu.com/news/international
[8]http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words
[9]We removed the stop words after extracting the opinion sentences in Section 4.2 because some stop words are indicative opinion clues such as "should" and "must"

and adverbs that only appear in the opinion sentences, because these words are more likely to convey the opinions. To judge whether a sentence expresses an opinion or not, we choose the opinion clues as basic criteria. Opinion clues are used in [7] to extract opinion sentences from blog pages and are also used in [3] to extract statements which best express opinionists' standpoints on certain topics. Following these work, we use the rule-based method to define opinion clues. More details can be found in [7, 3]. In addition, we also augment the opinion clues by adding their synonyms through WordNet[10] and those opinion words included in MPQA Opinion Corpus[11] [33]. To classify tokens into nouns, adjectives, verbs and adverbs, we use the Part-of-Speech tagging function provided by the MontyLingua Python library[12].

4.3 Research Questions

An extensive set of experiments are designed to address the following questions of the proposed research:

- Can the CPT model effectively discover the shared topics across multiple perspectives and accurately capture the opinions expressed by different perspectives on the topics? (Section 5.2)

- Can CPT have improved predictive performance over other methods for opinion modeling? (Section 5.1.1)

- Can CPT effectively present opinion words for ad hoc queries? (Section 5.1.2)

- Can the diversity metric derived from the learned model parameters characterize the difference of multiple perspectives? (Section 5.2.1)

5. EXPERIMENTS

In this section we present both quantitative and qualitative experiments on the two testbeds. For the quantitative evaluation, we show that CPT performs substantially better than the baseline methods. For the qualitative analysis we show that the opinions inferred by CPT do accurately correspond to their perspectives on the topics.

5.1 Quantitative Evaluation

5.1.1 Opinion Perplexity

In this experiment, we use perplexity as the criterion for model evaluation. Perplexity is a quantitative measure for comparing language models and is often used to compare the predictive performance of topic models [8]. The value of perplexity reflects the ability of a model to generalize to unseen data. A lower perplexity score indicates better generalization performance. In our case, perplexity reflects the ability of a model to predict opinion words for new unseen documents. The perplexity is algebraically equivalent to the inverse of the geometric mean of per-word (per-opinion word in our case) likelihood. Formally, the perplexity for a set of test documents D_{test} is calculated as follows:

$$perplexity(D_{test}) = \exp -\frac{\sum_{d=1}^{|D_{test}|} \log\big(p(\boldsymbol{o}_d)\big)}{\sum_{d=1}^{|D_{test}|} N_{o,d}} \quad (2)$$

---

[10]http://wordnet.princeton.edu
[11]http://www.cs.pitt.edu/mpqa/databaserelease
[12]http://web.media.mit.edu/hugo/montylingua/index.html

Table 2: Statistics of the Senate and News testbeds

	Senate			News			
	Republican	Democrat	Total	NYT	Xinhua	Hindu	Total
Number of documents	4,097	9,876	13,973	8,225	4,177	3,731	16,133
Number of sentences	137,688	285,804	423,492	219,766	48,111	57,513	325,390
Number of words	3,358,239	7,340,255	10,698,494	5,753,693	1,715,817	599,222	7,868,732
Number of topic words	697,003	1,546,911	2,243,914	1,185,518	396,464	125,804	1,707,786
Number of opinion words	768,367	347,709	1,116,076	573,560	200,546	158,176	932,282

Table 3: 20 ad hoc queries for each testbed

Senate	News
immigration, Iraq war, abortion, healthcare, education, veteran, agriculture, censorship, drugs, taxes, stem cell, minimum wage, trade, financial market, climate change, guns, death penalty, judges, prayer, affirmative action	Dalai Lama, Kashmir, Wikileaks, nuclear weapon, iphone, climate change, terrorism, Haiti earthquake, Iran, WTO, Xiaobo Liu, Islam, corruption, Google, energy, communist, education, censorship, population control, globalization

where

$$p(\boldsymbol{o}_d) = \prod_{i=1}^{N_{o,d}} \sum_{k=1}^{K} p(o_i|z_i = k)p(z_i = k|d) \qquad (3)$$

In the above equation $\boldsymbol{o}_d$ is the set of opinion words appearing in the test document d. The probability $p(o_i|z_i = k)$ is learned from the training process, and $p(z_i = k|d)$ is inferred from a Gibbs Sampling process on the test data based on the parameters learned from training data. We randomly select 20% of the documents as a held-out test data and train the model on the remaining 80%.

In topic models, we need to select the number of topics. A range of 50 to 300 topics is typically used in the literature. 50 topics are often used for relatively small collections and 300 for large collections. We test the perplexity of the trained model on the test data for different topic numbers K. Figure 2 shows the perplexities in five different settings of K for the Senate dataset. We can see that in general the perplexity scores for all the settings decrease over the iterations. The algorithm tends to converge after about 100 iterations. Along the iterations, larger topic number usually leads to smaller perplexity value, indicating a better prediction performance. This is due to the fact that the increased topic number reduces the uncertainty in training. The effect of increase in topic number on perplexity value gets smaller when the topic number gets larger. When the topic number set to 160, the perplexity value increases. Therefore, we set the topic number $K = 120$ which leads to a near-minimum perplexity. The topic number selection process for the News dataset is similar and the results are also shown in Figure 2.

In this subsection, we compare CPT with LDA and corrLDA on the metric of perplexity. In the experiments, we adapt corrLDA to opinion modeling by changing the Gaussian distribution over image features to multinomial distribution over topic words. The Gibbs sampling procedure can be similarly derived. As discussed in Section 3, in LDA, topic words and opinion words are mixed together and generated from a single distribution. Therefore, it is not an appropriate comparison with CPT if we use $p(o)$ in LDA to calculate the perplexity (because $\sum_o p(o) < 1$). Instead, we train a LDA model only on the opinion words, which make it use the same opinion word vocabulary with CPT.

Figure 3 plots the perplexity results for each model over different topic numbers. The iteration numbers for both models are set to 120. From the plots, we can see that CPT achieves the minimum perplexity on both testbeds among the three models. When the number of topics is small, CPT and corrLDA yield similar performance. When K gets large, the gap between CPT and corrLDA is generally widened. These results may be explained by the fact that corrLDA does not differentiate the perspectives of documents while CPT does. In consequence, when finer granularity of topics is present, CPT can yield better predicative performance than corrLDA because opinions for a more focused topic are generally more homogeneous. Furthermore, CPT and corrLDA seem less affected by the topic number. In fact, CPT and corrLDA shows consistent performance in a wide range of topic numbers from 100 to 200. In addition, on both datasets, LDA achieves its minimum with a smaller topic number than CPT and corrLDA. This may be explained by the fact the topics in LDA derived from the opinion words while the topics in CPT and corrLDA come from the topic words. There are more topic words than opinion words (as shown in Table 2), which probably results in more heterogeneous topics in CPT and corrLDA.

5.1.2 Ad hoc Queries

In this subsection, we conduct quantitative experiments to evaluate the retrieval performance of CPT on ad hoc query topics. Table 3 shows the 20 ad hoc queries for each testbed. These queries are chosen based on several knowledge sources[13] and they are perceived to have varied degree of different stances among the perspectives. When selecting these queries, we did not know what topics would be learned from the model. For each query, the model returns a ranked list of opinion words for each perspective. The results are judged by two people who are familiar with the query topics. A binary judgment (i.e. "relevant" or "not relevant") is made for each opinion word against the query topic. The evaluation metrics are Precision at 5 (P@5), Precision at 10 (P@10) Precision at 20 (P@20), Mean Reciprocal Rank (MRR), and normalized Discounted Cumulative Gain at 20 (nDCG@20).

In Table 4, we compare the models with and without ap-

---

[13]http://www.americanpolitics.com/030499dictionary.html
http://www.diffen.com/difference/Democrat_vs_Republican
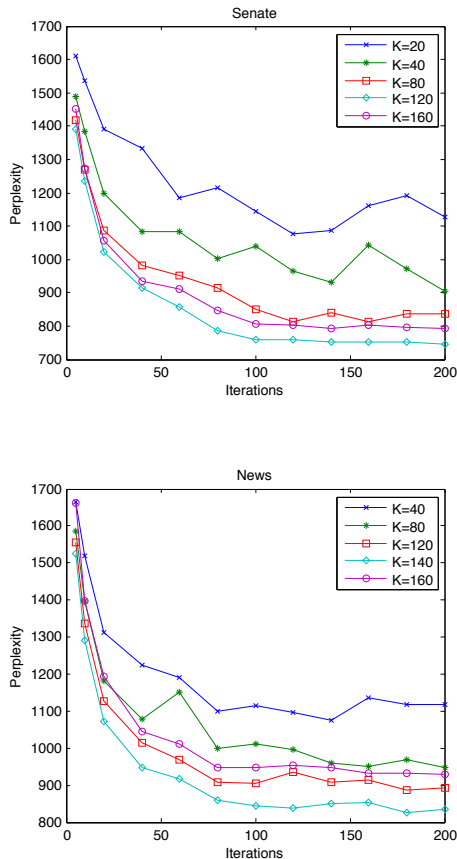http://www.cfr.org/india/india-china-united-states-delicate-balance/p9962

Figure 2: The perplexity results of the Cross-Perspective Topic model on the two testbeds for six different topic numbers K over the iterations. Top: Senate dataset. Bottom: News dataset.

Figure 3: The perplexity results of LDA, corrLDA and CPT on the two testbeds. Left: Senate dataset with topic number K=20, 40, 80, 120, 160 and 200. Right: News dataset with topic number K=40, 80, 120, 140, 160 and 200

Table 4: Evaluation results of CPT with and without extracting opinion sentences. Best results on each testbed are highlighted. The †symbol indicates statistical significance (by two-tailed Student's t-test) of "FULL" against "OS" at 0.95 confidence interval.

	P@5	P@10	P@20	MRR	nDCG
Senate					
FULL	0.835	0.786	0.688	0.911	0.773
OS	0.896†	0.824	0.727	0.952†	0.822†
News					
FULL	0.764	0.745	0.639	0.875	0.714
OS	0.822†	0.782	0.674	0.901	0.768†

plying the opinion word extraction procedure presented in Section 4.2. "FULL" represents the opinion words come from the whole document. "OS" represents the opinion words only come from the extracted opinion sentences. By comparing "OS" with "FULL", we can see quite a big positive impact from our opinion word extraction method.

To compare CPT with other methods, we use LDA and corrLDA as two baselines. Specifically, we train LDA and corrLDA models on the whole collection of each perspective and then calculate $p(w|q)$ in a similar way as shown in Eqn. (1). Because the words in LDA contain both topic and opinion words, we only focus on the opinion words and rank them according to $p(w|q)$. In addition, we design another two retrieval based baselines for comparison as follows. For each query topic, we retrieve 50 documents from each perspective by BM25 [21]. We then extract the opinion words using the procedure in Section 4.2. For the first baseline ("freq"), we rank the opinion words by frequency in each perspective. For the second baseline ("mutual"), we rank the opinion words by mutual information with respect to each perspective. Mutual information measures how much information (in the information-theoretic sense) an opinion word contains about the perspective [21]. Table 5 presents the comparison of the five methods.

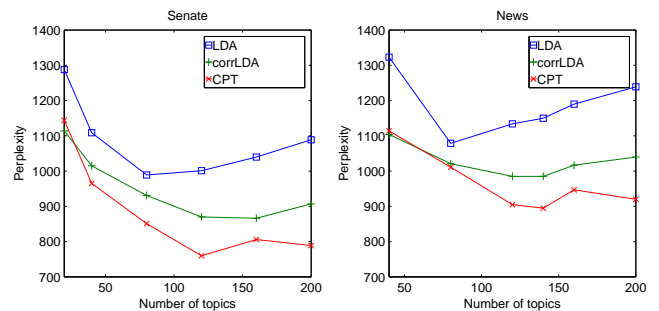From the table, we can see that CPT achieves the best

results on both testbeds in all the evaluation metrics. By comparing CPT with the baselines, we can see that CPT has substantial improvement especially on P@5, MRR and nDCG@20, which indicates CPT is effective to return the relevant results to the top of the list. corrLDA generates the second best results after CPT, while there exists noticeable gaps between these two methods. For the retrieval based method, "mutual" generally yields better results than "freq", and "freq" yields similar performance than LDA on both testbeds. Another observation is that the performance on the News dataset is worse than on the Senate dataset. This can be explained by the fact that Republican and Democratic senators usually have shared issues to discuss which enables the models to learn more representative topics and more logically connected opinions. In contrast, the news coverage of the three news agencies could be more diverse. This observation is also consistent with the perplexity results in Figure 3.

5.2 Qualitative Analysis

In this subsection, we show the discovered topics and the corresponding opinions by the Cross-Perspective Topic model. The model is trained on the whole collections with the parameters chosen based on the experimental results in Section 5.1.1. Table 6 contains a sample of topics and the corresponding opinions on the Senate dataset. The top 5 words from the shared topic-word distribution $p(w|z)$ and the top 5 words from the topic-opinion distribution $p(o|z)$ are shown for each perspective. The coupling between $p(w|z)$ and $p(o|z)$ can illustrate each perspective's opinions o on the topic z represented by the word w. By looking at the ta-

Table 6: A sample of topics and the corresponding opinions from Republican and Democratic parties. Shown are the top 5 words from the shared topic-word distribution $p(w|z)$ and the top 5 words from the topic-opinion distribution $p(o|z)$ for each party.

		Republican		Democrat	
TOPIC 9					
Word	**Prob.**	**Opinion**	**Prob.**	**Opinion**	**Prob.**
immigration	0.1165	illegal	0.0370	comprehensive	0.0330
border	0.0761	alien	0.0362	legal	0.0275
reform	0.0415	secure	0.0317	fair	0.0251
security	0.0285	comprehensive	0.0290	undocumented	0.0204
visa	0.0277	enforce	0.0286	temporary	0.0202
TOPIC 26					
insurance	0.1255	small	0.0344	uninsured	0.0393
health	0.1227	private	0.0198	federal	0.0281
coverage	0.0732	eligible	0.0181	affordable	0.0205
care	0.0394	responsible	0.0177	expand	0.0187
medicaid	0.0358	individual	0.0175	public	0.0185
TOPIC 39					
Word	**Prob.**	**Opinion**	**Prob.**	**Opinion**	**Prob.**
trade	0.1449	global	0.0195	domestic	0.0198
agreement	0.0604	developing	0.0190	unfair	0.0187
china	0.0331	unfair	0.0163	lost	0.0171
manufacturing	0.0255	manipulate	0.0161	fair	0.0169
world	0.0179	competitive	0.0160	environmental	0.0146
TOPIC 75					
Word	**Prob.**	**Opinion**	**Prob.**	**Opinion**	**Prob.**
iraq	0.1692	military	0.0296	military	0.0177
war	0.0614	supplemental	0.0156	failed	0.0164
security	0.0189	win	0.0151	end	0.0157
afghanistan	0.0171	critical	0.0147	change	0.0147
saddam	0.0170	secure	0.0147	withdraw	0.0145

Table 5: Comparison of CPT with other methods for each testbed. Best results on each testbed are highlighted. The †symbol indicates statistical significance (by two-tailed Student's t-test) of "CPT" against "freq" at 0.95 confidence interval.

	P@5	P@10	P@20	MRR	nDCG
Senate					
freq	0.818	0.769	0.672	0.875	0.771
mutual	0.832	0.788	0.685	0.896	0.778
LDA	0.812	0.764	0.672	0.882	0.759
corrLDA	0.859†	0.798	0.701	0.922†	0.802
CPT	**0.896†**	**0.824†**	**0.727**	**0.952†**	**0.822†**
News					
freq	0.758	0.740	0.638	0.853	0.711
mutual	0.767	0.748	0.645	0.879	0.721
LDA	0.751	0.738	0.642	0.857	0.717
corrLDA	0.792	0.766	0.659	0.886†	0.730
CPT	**0.822†**	**0.782**	**0.674**	**0.901†**	**0.768†**

ble we can see some clear differences between Republicans and Democrats. For example, in Topic 9 which is about immigration, Republican senators often used the words "illegal aliens" while Democratic senators probably prefer to use "undocumented" workers/immigrants. In fact, the choice of words can reflect their stance on the issue[14]. In addition, Republicans seem to emphasize more on the "secure" aspect of the immigration reform while Democrats more on the "legal" and "fair" aspect. In Topic 26 which is about health insurance, "private" and "individual" vs "public" and "federal" probably indicates the two parties' difference on the

role of government in this issue. Topic 39 is about the trade with China and both parties seem to think it is "unfair" while the other opinion words are different. In Topic 75 about Iraq war, "win" vs "failed" and "supplemental" vs "withdraw" also illustrate their different attitudes towards the war.

Table 7 presents a sample of topics and the corresponding opinions from the News dataset. From the table, it is interesting to see the media bias on the issues. For example, topic 40 is about 2010 Nobel Peace Prize laureate Liu Xiaobo (actually the 7th top word is "xiaobo" which is not shown in the table). In this topic, we can clearly see the huge discrepancy between New York Times (The Hindu) and Xinhua on this issue. The difference is also manifested in Topic 54 which is about Iran uranium enrichment. New York Times probably reports more on potential "military" operations while Xinhua emphasizes on "peaceful", "diplomatic" and "negotiate" aspects. Hindu seems to have a middle ground perspective between NYT and Xinhua. In Topic 68 which is about Kashmir in Pakistan, all the top 5 opinion words from the three news agencies are different, which may indicate their different views on this issue. In Topic 81 which is about China, although "economic" is the top word in all the three perspectives, the other opinion words clearly show their differences especially between Xinhua and the other two media. For example, the American and India media often use the word "rising" while the Chinese media uses "developing". This can be explained by the fact that "China's peaceful rise" was replaced in Chinese government parlance from 2004 by "China's peaceful development", to emphasize that China poses no threat to the established order[15].

5.2.1 Diversity Metric

---

[14]The term "illegal aliens" is considered offensive to some Latinos: https://www.spj.org/quill_issue.asp?ref=1745

[15]http://en.wikipedia.org/wiki/China's_peaceful_rise

Table 7: A sample of topics and the corresponding opinions from the three media: New York Times in US, Xinhua News in China and The Hindu in India. Shown are the top 5 words from the shared topic-word distribution $p(w|z)$ and the top 5 words from the topic-opinion distribution $p(o|z)$ for each news agency.

		New York Times		Xinhua News		The Hindu	
TOPIC 40							
WORD	**PROB.**	**OPINION**	**PROB.**	**OPINION**	**PROB.**	**OPINION**	**PROB.**
peace	0.0573	dissident	0.0116	convicted	0.0163	jailed	0.0168
prize	0.0533	awarded	0.0101	arrogant	0.0114	dissident	0.0077
nobel	0.0425	democratic	0.0078	political	0.0092	criticised	0.0062
liu	0.0391	imprisoned	0.0068	interfere	0.0092	unaware	0.0062
committee	0.0281	pro-democracy	0.0049	internal	0.0085	imprisoned	0.0046
TOPIC 54							
WORD	**PROB.**	**OPINION**	**PROB.**	**OPINION**	**PROB.**	**OPINION**	**PROB.**
iran	0.2209	military	0.0765	peaceful	0.1065	diplomatic	0.0713
program	0.0391	impose	0.0623	diplomatic	0.0128	international	0.0496
tehran	0.0334	stop	0.0442	negotiate	0.0121	military	0.0367
uranium	0.0305	diplomatic	0.0307	civilian	0.0113	constructive	0.0299
ahmadinejad	0.0195	financial	0.0225	unilateral	0.0109	regional	0.0186
TOPIC 68							
WORD	**PROB.**	**OPINION**	**PROB.**	**OPINION**	**PROB.**	**OPINION**	**PROB.**
kashmir	0.0404	ethnic	0.0147	indian-controlled	0.0205	democratic	0.0275
pakistan	0.0388	killed	0.0147	moderate	0.0198	civil	0.0215
constitution	0.0222	disputed	0.0145	end	0.0162	insurgent	0.0201
violence	0.0193	peaceful	0.0145	infiltrate	0.0156	separate	0.0195
valley	0.0157	tibetan	0.0142	bilateral	0.0151	military	0.0184
TOPIC 81							
WORD	**PROB.**	**OPINION**	**PROB.**	**OPINION**	**PROB.**	**OPINION**	**PROB.**
china	0.2414	economic	0.1133	economic	0.1175	economic	0.1147
chinese	0.1063	state-run	0.0165	western	0.0520	communist	0.0193
beijing	0.0672	rising	0.0159	peaceful	0.0180	growing	0.0186
government	0.0213	manipulate	0.0156	positive	0.0171	rising	0.0180
currency	0.0109	controlled	0.0154	developing	0.0160	territorial	0.0179

Figure 4: Quantitative differences between Republican senators and Democratic senators on 8 ad hoc queries

Figure 5: Quantitative differences among New York Times, Xinhua News and The Hindu on 8 ad hoc queries

In this subsection, we use the diversity metric based on the Jensen-Shannon divergence (defined in Section 3.2) to quantify the dissimilarities between different perspectives with respect to various query topics. Due to space constraints, we only present 8 queries for each testbed. These queries are ad hoc as well, although some of them may correspond to the conceptual topics learned from the model. Figure 4 shows the results for the Senate dataset. From Figure 4, we
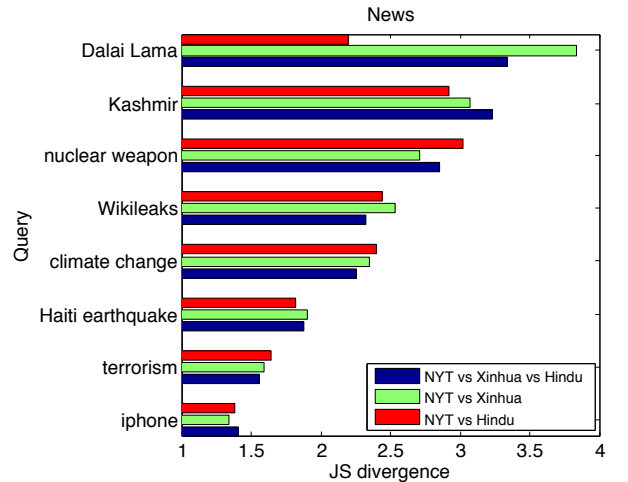
can see that the Republic and Democratic parties have quite different stances on the queries "immigration", "Iraq war", "abortion" and "healthcare". On the other hand, two parties have quite similar stances on "censorship", "agriculture" and "veteran". With respect to the query "education", the two parties have mild differences. These findings are consistent with what are commonly perceived about the two parties. Figure 5 shows the results for the News dataset. Besides showing the Jensen-Shannon divergence among the

three news agencies (i.e., NYT vs Xinhua vs Hindu), we also show the JS divergence between NYT and Xinhua, and the divergence between NYT and Hindu. Overall, we can see that the three agencies have very different opinions on the queries "Dalai Lama", "Kashmir", and "nuclear weapon". They have quite similar opinions on "iphone", "terrorism" and "Haiti earthquake", and have mild difference on "Wikileaks" and "climate change". By looking at the pair comparison, we can see that some big JS divergences are caused by pairwise difference. For example, NYT and Hindu actually have similar opinions on "Dalai Lama", while the JS divergence of the three agencies on the topic is the largest. On the other hand, on the query topic "nuclear weapon", NYT and Xihuna have more similar opinions. From Figure 5, we can see that generally NYT and Hindu share more similar stances on the issues while NYT and Xinhua hold more different opinions. From Figure 4 and Figure 5, we can easily identify the controversial and consensual issues among different perspectives. They could be a visualization tool to help reduce users' cognitive efforts.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we study a novel opining mining problem referred to as contrastive opining modeling. The work reported in this paper is just an initial step towards a promising new direction. There are many interesting future research directions. It is worth exploring the applicability of the CPT model to large-scale Web documents such as blogs and reviews. It is also interesting to see the proposed model to serve as a data mining tool for comparative research in social science such as cross-culture, cross-religion, and cross-country studies.

7. ACKNOWLEDGMENTS

8. REFERENCES

[1] D. Blei and M. Jordan. Modeling annotated data. In *SIGIR*, pages 127–134. ACM, 2003.
[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.
[3] B. Chen, L. Zhu, D. Kifer, and D. Lee. What is an opinion about? exploring political standpoints using opinion scoring model. In *AAAI*, pages 1007–1012.
[4] D. Diermeier, J. Godbout, B. Yu, and S. Kaufmann. Language and ideology in Congress. In *Annual Meeting of the Midwest Political Science Association*, 2007.
[5] X. Ding, B. Liu, and P. Yu. A holistic lexicon-based approach to opinion mining. In *WSDM*, pages 231–240, 2008.
[6] K. Eguchi and V. Lavrenko. Sentiment retrieval using generative models. In *EMNLP*, pages 345–354, 2006.
[7] O. Furuse, N. Hiroshima, S. Yamada, and R. Kataoka. Opinion sentence search engine on open-domain blog. In *IJCAI*, pages 2760–2765, 2007.
[8] T. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101:5228, 2004.
[9] V. Hatzivassiloglou and K. McKeown. Predicting the semantic orientation of adjectives. In *EACL*, pages 174–181, 1997.
[10] M. Hu and B. Liu. Mining opinion features in customer reviews. In *AAAI*, pages 755–760, 2004.
[11] M. Hurst and K. Nigam. Retrieving topical sentiments from online document collections. In *Document Recognition and Retrieval XI*, pages 27–34, 2004.
[12] Y. Jo and A. Oh. Aspect and sentiment unification model for online review analysis. In *WSDM*, pages 815–824, 2011.
[13] S. Kim and E. Hovy. Determining the sentiment of opinions. In *COLING*, pages 1367–1374, 2004.
[14] K. Lerman and R. McDonald. Contrastive summarization: an experiment with consumer reviews. In *NAACL/HLT*, pages 113–116, 2009.
[15] S. Li, C. Lin, Y. Song, and Z. Li. Comparable entity mining from comparative questions. In *ACL*, pages 650–658. Association for Computational Linguistics, 2010.
[16] C. Lin and Y. He. Joint sentiment/topic model for sentiment analysis. In *CIKM*, pages 375–384. ACM, 2009.
[17] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 2002.
[18] W. Lin and A. Hauptmann. Are these documents written from different perspectives?: a test of different perspectives based on statistical distribution divergence. In *ACL*, pages 1057–1064, 2006.
[19] W. Lin, T. Wilson, J. Wiebe, and A. Hauptmann. Which side are you on?: identifying perspectives at the document and sentence levels. In *CoNLL*, pages 109–116. Association for Computational Linguistics, 2006.
[20] B. Liu, M. Hu, and J. Cheng. Opinion observer: Analyzing and comparing opinions on the web. In *WWW*, pages 342–351, 2005.
[21] C. Manning, P. Raghavan, and H. Schutze. *Introduction to information retrieval*. Cambridge University Press, UK, 2008.
[22] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *WWW*, pages 171–180, 2007.
[23] I. Ounis, M. Rijke, C. Macdonald, G. Mishne, and I. Soboroff. Overview of the TREC-2006 Blog Track. In *TREC*, pages 15–27, 2006.
[24] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.
[25] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP*, pages 79–86, 2002.
[26] M. Paul and R. Girju. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *EMNLP*, pages 1408–1417, 2009.
[27] M. Paul, C. Zhai, and R. Girju. Summarizing contrastive viewpoints in opinionated text. In *EMNLP*, pages 66–76, 2010.
[28] A. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *HLT/EMNLP*, pages 339–346, 2005.
[29] Y. Seki, D. Evans, L. Ku, H. Chen, N. Kando, and C. Lin. Overview of opinion analysis pilot task at NTCIR-6. In *NTCIR-6*, pages 265–278, 2007.
[30] M. Thomas, B. Pang, and L. Lee. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *EMNLP*, pages 327–335, 2006.
[31] I. Titov and R. McDonald. A joint model of text and aspect ratings for sentiment summarization. In *ACL*, page 308, 2008.
[32] I. Titov and R. McDonald. Modeling online reviews with multi-grain topic models. In *WWW*, pages 111–120, 2008.
[33] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. OpinionFinder: A system for subjectivity analysis. In *HLT/EMNLP*, pages 34–35, 2005.
[34] T. Yano, W. Cohen, and N. Smith. Predicting response to political blog posts with topic models. In *NAACL/HLT*, pages 477–485, 2009.
[35] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *ICDM*, pages 427–434. IEEE, 2003.
[36] C. Zhai, A. Velivelli, and B. Yu. A cross-collection mixture model for comparative text mining. In *SIGKDD*, pages 743–748, 2004.
[37] M. Zhang and X. Ye. A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval. In *SIGIR*, pages 411–418. ACM, 2008.
[38] W. Zhang, C. Yu, and W. Meng. Opinion retrieval from blogs. In *CIKM*, pages 831–840. ACM, 2007.