Learning Topics and Positions from Debatepedia

Swapna Gottipati[†] **Minghui Qiu**[†] **Yanchuan Sim**[‡] **Jing Jiang**[†] **Noah A. Smith**[‡]
[†]School of Information Systems, Singapore Management University, Singapore
[‡]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[†]{swapnag.2010,minghui.qiu.2010,jingjiang}@smu.edu.sg
[‡]{ysim,nasmith}@cs.cmu.edu

Abstract

We explore Debatepedia, a community-authored encyclopedia of sociopolitical debates, as evidence for inferring a low-dimensional, human-interpretable representation in the domain of issues and positions. We introduce a generative model positing latent topics and cross-cutting positions that gives special treatment to person mentions and opinion words. We evaluate the resulting representation's usefulness in attaching opinionated documents to arguments and its consistency with human judgments about positions.

1 Introduction

The social web has evolved into a forum for large portions of the population to discuss and debate complex issues of societal importance. Websites like Debatepedia,[1] an online, community-authored encyclopedia of debates (§2), seek to organize some of this exchange into structured information resources that summarize arguments and link externally to texts (editorials, blog posts, etc.) that express and evoke them. Empirical NLP, we propose, has a role to play in creating a more compact and easily-interpretable way to understand the opinion space. In particular, we envision applications to computational journalism, where there is high demand for transformation of and pattern discovery in unmanageable, unstructured, evolving data (including text) to inform the public (Cohen et al., 2011).

In this paper, we develop a generative model for discovering such a representation (§3), using Debatepedia as a corpus of evidence. We draw inspiration from Lin et al. (2008) and Ahmed and Xing (2010), who used generative models to infer **topics**—distributions over words—and other word-associated variables representing perspectives or ideologies. We view topics as lexicons, and propose that *grounding* a topic model with evidence beyond bags of words can lead to more lexicon-like representations. Specifically, our generative topic model grounds topics using the hierarchical organization of arguments within Debatepedia. Further, we use named entity recognition as a preprocessing step, an existing sentiment lexicon to construct an informed prior, and we incorporate a latent, discrete **position** variable that cuts across debates.[2]

We evaluate the model informally and formally (§4). Subjectively, the model identifies reasonable topic and perspective terms, and it associates topics sensibly with important public figures. In quantitative evaluations, we find the model's representation superior to topics from vanilla latent Dirichlet allocation (Blei et al., 2003) and the joint sentiment topic model (Lin and He, 2009) in matching external texts to debates. Further, the position variables can be used to infer the *side* of an argument within a debate; our model performs with an accuracy of 86% on position prediction of the debate argument. The cross-cutting position variable is not especially consistent with human judgments, suggesting that further knowledge sources may be required to improve interpretability across issues.

2 Data

Debatepedia, like Wikipedia, is constructed by volunteer contributors and has a system of community

[1] http://dbp.idebate.org

[2] This variable might serve to cluster debate sides according to "abstract beliefs commonly shared by a group of people," sometimes called *ideologies* (Van Dijk, 1998). We do not claim that our model infers ideologies (see §4).

Debate: *Gun control; should laws be passed to limit gun ownership further?*	
Question: *Self-defense – Is self-defense a good reason for gun ownership?*	
Side: Yes	Side: No
Argument: A citizen has a "right" to guns as a means to self-defense: Many groups argue that a citizen should have the "right" to defend themselves, and that a gun is frequently the ...	Argument: The protection of property is not a good justification for yielding a lethal weapon. While people have a right to their property, this should not justify wielding a lethal ...
Argument: Gun restrictions and bans disadvantage citizens against armed criminals. Citizens that are not allowed to carry guns are disadvantaged against lawless criminals that ...	Argument: Robert F. Drinan, Former Democratic US Congressman, "Gun Control: The Good Outweighs the Evil", 1976 – "These graphic examples of individual instances of ...
Question: *Economic benefits – Is gun control economically beneficial?*	
Side: Yes	Side: No
Argument: Lax gun control laws are economically costly. The Coalition for Gun Control claims that, "in Canada, the costs of firearms death and injury alone have been estimated at ...	Argument: Gun sports have economic benefits. Field sports bring money into poor rural economies and provide a motivation for landowners to value environmental protection.

Table 1: An example of a Debatepedia debate on the topic "Gun control."

moderation. Many of the debate issues covered are controversial and salient in current public discourse. Because it is primarily expressed as text, Debatepedia is a *corpus* of debate topics, but it is organized hierarchically, with multiple issues in each debate topic, questions within each issue, and arguments on two sides of each question. An important feature of the corpus is the widespread quotation and linking to external articles on the web, including news stories, blog postings, wiki pages, and social media forums; here we use these external articles in evaluation (§4).

Table 1 shows excerpts from a debate page[3] from Debatepedia. Each debate contains "questions," which reflect the different aspects of a debate. In this particular debate, there are 13 questions (2 shown), ranging from economic benefits to enforceability to social impacts. For each question, there are two distinct sides, each with its own set of supporting arguments. Many of these arguments also contains links to online articles where the quotes are extracted from (not shown in Table 1). For example, in the second argument on the "No" side, there is an inline link to the article written by Congressman Drinan.[4]

Within a debate topic, the sides cut across different questions, aligning arguments together. In general, the questions are phrased so that a consistent "pro" and "con" structure is apparent throughout each debate, aligned to a high-level question (i.e., the "Yes" sides of all the questions are consistent with the same side of the larger debate). The example of Table 1 deviates from this pattern, with the self-defense "Yes" arguing "no" to the high-level debate question—*Should laws be passed to limit gun ownership further?*—and the economic "Yes" arguing "yes" to the high-level question.

Debates	1,303
Arguments	33,556
Articles linked by exactly one argument	3,352
Tokens	1,710,814
Types (excluding NE mentions)	59,601
Person named entity mentions	9,496

Table 2: Debatepedia corpus statistics. Types and tokens include unigrams, bigrams and person named entities.

Table 2 presents statistics of our corpus.

2.1 Preprocessing

We scraped the Debatepedia website and extracted the debate, question, argument, and side structure of the debate topics. We crawled the external web articles that were linked from the Debatepedia arguments. For the web articles, we extracted the main text content (ignoring boilerplate elements such as navigation and advertisments) using Boil-

[3] http://dbp.idebate.org/en/index.php/Debate:_Gun_control

[4] http://www.saf.org/LawReviews/Drinan1.html

erpipe (Kohlschütter et al., 2010).[5] We tokenized the text and filtered stopwords.[6] We considered both unigrams and bigrams in our model, keeping all unigrams and removing bigram types that appeared less than 5 times in the corpus. Although our modeling approach ultimately treats texts as bags of terms (unigrams and bigrams), one important preprocessing step was taken to further improve the interpretability of the inferred representation: named entity mentions of persons. We identified these mentions of persons using Stanford NER (Finkel et al., 2005) and treated each person mention as a single token. In our qualitative analysis of the model (§4.2), we will show how this special treatment of person mentions enables the association of well-known individuals with debate topics. Though not part of our experimental evaluation in this paper, such associations are, we believe, an interesting direction for future applications of the model.

3 Model

Our model defines a probability distribution over terms[7] that are observed in the corpus. Each term occurs in a context defined by the tuple $\langle d, q, s, a \rangle$ (respectively, a *debate*, a *question* within the debate, a *side* within the debate, and an *argument*). At each level of the hierarchy is a different latent variable:

- Each question q within debate d is associated with a distribution over topics, denoted $\boldsymbol{\theta}_{d,q}$.[8]
- Each side s of the debate d is associated with a position, denoted $i_{d,s}$ and we posit a global distribution ι that cuts across different questions and arguments. In our experiments, there are two positions, and the two sides of a debate are constrained to associate with opposing positions. As illustrated by Table 1, this assumption is not always correct, though it tends to hold most of the time.
- Each term $w_{d,q,s,a,n}$ (n is the position index of the term within an argument) is associated with one of five *functional term types*, denoted $y_{d,q,s,a,n}$. This variable is latent, except when it takes the value "entity" (e) for terms marked as named entity mentions. When it is not an entity, it takes one of the other four values: "general position" (i), "topic-specific position" (o), "topic" (t), or "background" (b). Thus, every term w is drawn from one of these 5 types of bags, and y acts as a switching variable to select the type of bag.
- For some term types (the ones where $y \in \{\mathsf{o}, \mathsf{t}\}$), each term $w_{d,q,s,a,n}$ is associated with one of T discrete topics, as indexed by $z_{d,q,s,a,n}$.

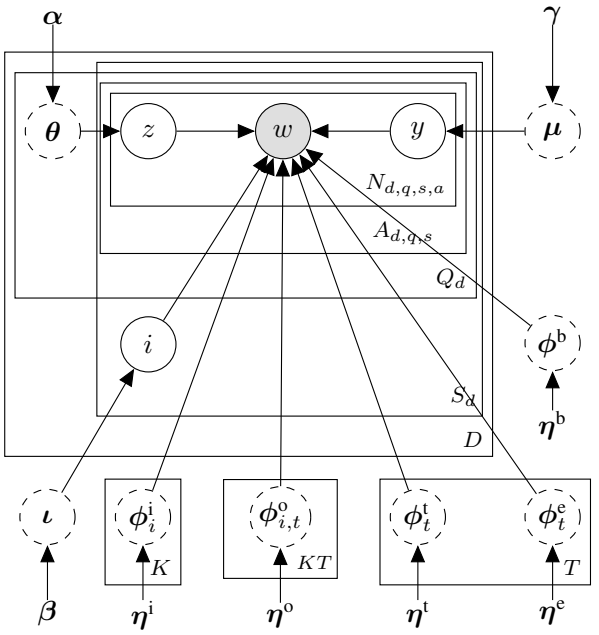

Figure 1: Plate diagram. K is the number of positions, and T is number of topics. The shaded variables are observed and dashed variables are marginalized. $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ and all $\boldsymbol{\eta}$ are fixed hyperparameters (§3.1).

Figure 1 illustrates the plate diagram for the graphical model underlying our approach. The generative story is given in Figure 2.

3.1 Priors

Typical probabilistic topic models assume a symmetric Dirichlet prior over its term distributions or

[5] http://code.google.com/p/boilerpipe

[6] www.ranks.nl/resources/stopwords.html

[7] Recall that our model includes bigrams. We treat each unigram and bigram token (after filtering discussed in §2.1) as a separate term.

[8] In future work, more sharing across questions within a debate, or more differentiation among the topic distributions for arguments under a question, might be explored. Wallach (2006) describes suitable techniques using hierarchical Dirichlet draws, and Eisenstein et al. (2011) suggests the use of sparse shocks to log-odds at different levels. Here we work on the assumption that Debatepedia's questions are the most topically coherent level, and work with a single topic mixture at this level.

1. $\forall$ topics t, draw topic-term distribution $\boldsymbol{\phi}^{\mathsf{t}}_t \sim \text{Dirichlet}(\boldsymbol{\eta}^{\mathsf{t}})$ and topic-entity distribution $\boldsymbol{\phi}^{\mathsf{e}}_t \sim \text{Dirichlet}(\boldsymbol{\eta}^{\mathsf{e}})$.
2. $\forall$ positions i, draw position-term distribution $\boldsymbol{\phi}^{\mathsf{i}}_i \sim \text{Dirichlet}(\boldsymbol{\eta}^{\mathsf{i}})$.
3. $\forall$ topics t, $\forall$ positions i, draw topic-position term distribution $\boldsymbol{\phi}^{\mathsf{o}}_{i,t} \sim \text{Dirichlet}(\boldsymbol{\eta}^{\mathsf{o}})$.
4. Draw background term distribution $\boldsymbol{\phi}^{\mathsf{b}} \sim \text{Dirichlet}(\boldsymbol{\eta}^{\mathsf{b}})$.
5. Draw functional term type distribution $\boldsymbol{\mu} \sim \text{Dirichlet}(\boldsymbol{\gamma})$.
6. Draw position distribution $\boldsymbol{\iota} \sim \text{Dirichlet}(\boldsymbol{\beta})$.
7. $\forall$ debates d:
   a. Draw $i_{d,1}, i_{d,2} \sim \text{Multinomial}(\boldsymbol{\iota})$, assigning each of the two sides to a position.
   b. $\forall$ questions q in d:
      i. Draw topic mixture proportions $\boldsymbol{\theta}_{d,q} \sim \text{Dirichlet}(\boldsymbol{\alpha})$.
      ii. $\forall$ arguments a under question q and term positions n in a:
         A. Draw topic label $z_{d,q,s,a} \sim \text{Multinomial}(\boldsymbol{\theta}_{d,q})$.
         B. Draw functional term type $y_{d,q,s,a} \sim \text{Multinomial}(\boldsymbol{\mu})$.
         C. Draw term $w_{d,q,s,a} \sim \text{Multinomial}\left(\boldsymbol{\phi}^{y_{d,q,s,a}} \mid i_{d,1}, i_{d,2}, z_{d,q,s,a}\right)$.

Figure 2: Generative story for our model of Debatepedia.

apply empirical Bayesian techniques to estimate the hyperparameters. Motivated by past efforts to exploit prior knowledge (Zhao et al., 2010; Lin and He, 2009), we use the OpinionFinder sentiment lexicon[9] (Wilson et al., 2005) to construct $\boldsymbol{\eta}^{\mathsf{i}}$ and $\boldsymbol{\eta}^{\mathsf{o}}$. Specifically, terms w in the lexicon were given parameters $\eta^{\mathsf{i}}_w = \eta^{\mathsf{o}}_w = 0.01$, and other terms were given $\eta^{\mathsf{i}}_w = \eta^{\mathsf{o}}_w = 0.001$, capturing our prior belief that opinion-expressing terms are likely to be used in expressing positions. 5,451 types were given a "boost" through this prior.

Information retrieval has long exploited the observation that a term's document frequency (i.e., the number of documents a term occurs in) is inversely related its usefulness in retrieval (Jones, 1972). We encode this in $\boldsymbol{\eta}^{\mathsf{b}}$, the prior over the background term distribution, by setting each value to the logarithm of the term's argument frequency.

The other priors were set to be symmetric: $\eta^{\mathsf{e}} = 0.01$ (entity topics), $\eta^{\mathsf{t}} = 0.001$ (topics), $\alpha = 50/T = 1.25$ (topic mixture coefficients), $\beta = 0.01$ (positions), and $\gamma = 0.01$ (functional term types). Preliminary tests showed that final topics are relatively insensitive to the values of the hyperparameters.

3.2 Inference and Parameter Estimation

Exact inference under this model, like most latent-variable topic models, is intractable. We apply collapsed Gibbs sampling, a standard approach for such models (Griffiths and Steyvers, 2004).[10] The notable deviations from typical uses of collapsed Gibbs sampling are: (i) we jointly sample $i_{d,1}$ and $i_{d,2}$ to respect the constraint that they differ; and (ii) we fix the priors, in some cases to be asymmetric, as discussed in §3.1. We perform Gibbs sampling for 2,000 iterations over the dataset, discarding the first 500 iterations for burn-in, and averaging over every 10th iteration thereafter to get estimates for our term distributions.

3.3 T and K

In all experiments, we use $T = 40$ topics and $K = 2$ positions. We did not extensively explore different values for T and K; preliminary exploration suggested that interpretability, gauged informally by the authors, degraded for higher values of either.

4 Evaluation

Recall that the aim of this work is to infer a low-dimensional representation of debate text. We estimated our model on the Debatepedia debates (not including hyperlinked articles), and conducted several evaluations of the model, each considering a different aspect of the goal. We exploit external articles hyperlinked from Debatepedia described in §2 as supporting texts for arguments, treating each one's association to an argument as variable to be predicted. Firstly, we evaluate our model on the article associating task. Secondly, we evaluate our model on the position prediction task. Then, we compare

[9] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

[10] Because this technique is well known in NLP, details are relegated to supplementary material.

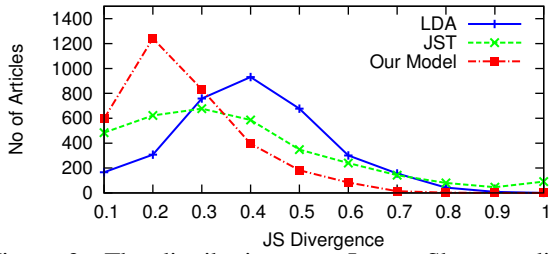

Figure 3: The distribution over Jensen-Shannon divergences between a hyperlinked article and the corresponding Debatepedia argument, $n = 3,352$.

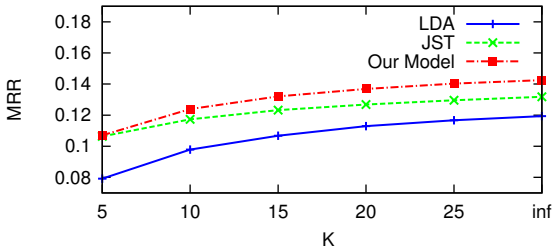

Figure 4: Mean reciprocal ranks for the association task.

our model's positional assignment of arguments to human annotated clusterings. Finally, we present qualitative discussion.

4.1 Quantitative Evaluation

4.1.1 Topics

As described in §2, our corpus includes 3,352 articles hyperlinked by Debatepedia arguments.[11] Our model can be used to infer the posterior over topics associated with such an article, and we compare that distribution to that of the Debatepedia article that links to it. Calculating the similarity of these distributions, we get an estimate of how closely our model can associate text related to a debate with the specific argument that linked to it. We compare with LDA (Blei et al., 2003), which ignores sentiment, and the joint sentiment topic (JST) model (Lin and He, 2009), an unsupervised model that jointly captures sentiment and topic.[12] Using Jensen-Shannon divergence, we find that our approach embeds these pairs significantly closer than LDA and JST (also trained with 40 topics), under a Wilcoxon signed rank test ($p < 0.001$). Figure 3 shows the histogram of divergences between our model, JST, and LDA.

Associating external articles. More challenging, of course, is *selecting* the argument to which an external article should be associated. We used the Jensen-Shannon divergence between topic distributions of articles and arguments to rank the latter, for each article. The mean reciprocal rank scores (Voorhees, 1999) for LDA, JST, and our model were 0.1272, 0.1421, and 0.1507, respectively; the difference is significant (Wilcoxon signed rank test, $p < 0.001$). We found the same pattern for MRR@k, $k \in \{5, 10, 15, 20, 25, \infty\}$, as shown in Figure 4.

It is likely possible to engineer more accurate models for attaching articles to arguments, but the attachment task is our aim only insofar as it contributes to an overall assessment of an inferred representation's quality.

4.1.2 Positions

Positional distance by topic. We next consider the JS divergences of position term distributions by topic; for each topic t, we consider the divergence between inferred values for $\phi^{\circ}_{1,t}$ and $\phi^{\circ}_{2,t}$. Figure 5 shows these measurements sorted from most to least different; these might be taken as evidence for which issue areas' arguments are more *lexically* distinguishable by side, perhaps indicating less common ground in discourse or (more speculatively) greater controversy. For example, our model suggests that debates relating to topics like presidential politics, foreign policy, teachers, women's health, religion, and Israel/Palestine are more heated (within the Debatepedia community at the time the debates took place) than those about the minimum wage, Iran as a nuclear threat, or immigration.

Predicting positions for arguments. We tested our model's ability to infer the positions of arguments. In this experiment (only), we held out 3,000 arguments during parameter estimation. The held-out arguments were selected so that every debate side maintained at least one argument whose inferred side could serve as the correct answer for the held-out argument. We then inferred i for each held-out argument from debate d and side s, given the parameters, and compared it with the value of $i_{d,s}$ inferred during parameter estimation. The model achieved 86% accuracy (Table 3 shows the confu-

[11] We consider only those articles linked by a single Debatepedia argument.

[12] JST multiplies topics out by the set of sentiment labels, assigning each token to both a topic and a sentment. We use the OpinionFinder lexicon in JST's prior in the same way it is used in our model.

sion matrix). Note that JST does not provide a baseline for comparison, since it does not capture debate sides.

	$i=1$	$i=2$
$i^*=1$	1,272	216
$i^*=2$	199	1,313

Table 3: Confusion matrix for position prediction on held-out *arguments*.

Predicting positions for external articles. We can also use the model to predict the position adopted in an external text. For articles linked from within Debatepedia, we have a gold standard: from which side of a debate was it linked? After using the model to infer a position variable for such a text, we can check whether the inferred position variable matches that of the argument that links to it. Table 4 shows that our model does not successfully complete this task, assigning about 60% of both kinds of articles $i = 1$.

	$i=1$	$i=2$
$i^*=1$	1,042	623
$i^*=2$	1,043	644

Table 4: Confusion matrix for position prediction on hyperlinked *articles*.

Genre. We manually labeled 500 of these articles into six genre categories. We had two annotators for this task (Cohen's $\kappa = 0.856$). These categories, in increasing order of average Jensen-Shannon divergence, are: blogs, editorials, wiki pages, news, other, and government. Figure 6 shows the results. While the only difference between the first and last groups are surprising by chance, we are encouraged by our model's suggestion that blogs and editorials may be more "Debatepedia argument-like" than news and government articles.

Note that our model is learned only from text *within* Debatepedia; it does not observe the text of external linked articles. Future work might incorporate this text as additional evidence in order to capture effects on language stemming from the interaction of position and genre.

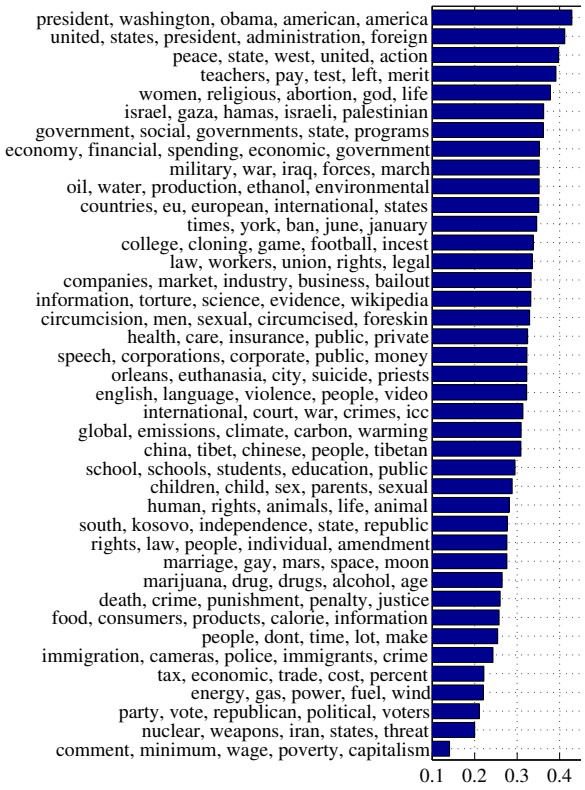

Figure 5: Jensen-Shannon divergences between topic-specific positional term distributions, for each topic. Topics are labeled by their most frequent terms from ϕ^{t}.

4.1.3 Comparison to Human Judgments of Positions

We compared our model's inferred positions to human judgments. For each of the 11 topics in Table 8, we selected two associated debates with more arguments than average (24.99). The debates were provided to each of three human annotators,[13] who

[13] All were native English-speaking American graduate students not otherwise involved in this research. Each is known by the authors to have basic literacy with issues and debates in

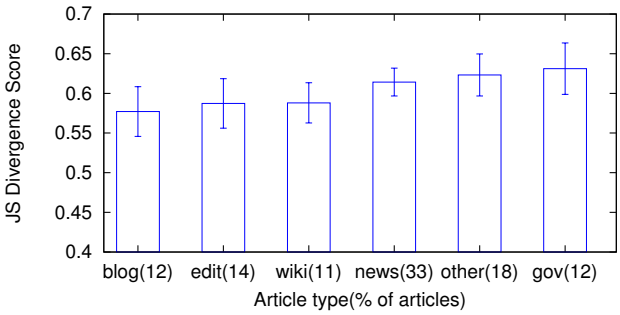

Figure 6: Position prediction on 500 hyperlinked *articles* by genre.

	"Israel-Palestine"	"Same-sex marriage"	"Drugs"	"Healthcare"	"Death penalty"	"Abortion"
i_1	pre emptive	same sex	hands free	single payer	anti death	pro choice
	israeli palestinian	long term	performance enhancing	so called	non violent	pro life
	open and shut	second class	in depth	self sustaining	african american	non muslim
i_2	two state	opposite sex	long term	government run	semi automatic	would be
	long term	well intentioned	high speed	government approved	high profile	full time
	self destructive	day time	short term	high risk	hate crime	late term

a. Our model: topic-specific position bigrams associated with six selected topics.

–	war	large	illegal	support	death	power
	assault	possibility	abuse	force	penalty	limit
	disproportionate	problems	high	threat	murder	civil
+	peace	civil	disease	care	power	care
	independence	rights	nature	universal	clean	suicide
	self-determination	affirmative	potential	uninsured	waste	death

b. JST: sentiments associated with six selected topics manually aligned to our model's topics.

Table 6: Terms associated with selected topics. The labels and alignments between the two models' topics were assigned manually. (a.) Our model: topic-specific position bigrams which are ranked by comparing the log odds conditioned on the position and topic: $\log \phi^{o}_{i_1,t,w} - \log \phi^{o}_{i_2,t,w}$. We show the top three terms for each position (b.) JST: we show the top three terms for each sentiment (negative and positive).

	A1 (11)	A2 (5)	A3 (16)
Model (2)	3.21	2.58	3.45
A1 (11)		2.15	2.15
A2 (5)			2.63

Table 5: Variation of information scores for each pairing of annotators and model.

were instructed to group the 44 sides of the debates. The instructions stated:

> Our goal is to see what you think about how the different sides of different debates can be lined up. You might find it convenient to think of these in terms of political philosophies, contemporary political party platforms, or something else. Any of these is fine; we want you to tell us the grouping you find most reasonable.

All three annotators (hereafter denoted A1, A2, and A3) used fairly involved labeling schemes; the annotators used 37, 30, and 16 unique labels, respectively.[14] A1 used keyword lists to label items; we coarsened his labels manually by removing or merging less common keywords (resulting in: *Republican*, *Democrat*, *science/environment*, *nanny*, *political reform*, *fiscal liberal*, *fiscal conservative*, *libertarian*, *Israel*, *Palestine*, and one unlabeled side). A2 provided a coarse annotation along with each fine-grained one (*liberal*, *conservative*, *?*, and two unlabeled sides). We used 100 samples from our Gibbs sampler to estimate posteriors for each $i_{d,s}$; these were always 99% or more in agreement, so we mapped each debate side into its single most probable cluster. Recall that the two sides of each debate must be in different clusters.

Table 5 shows the variation of information measure (Meila, 2003) for each pairing among the three annotators and our model. The model agrees with A2's coarse clustering most closely, and in fact is closer to A2's clustering than A2 is to A3's; it also agrees with A2's coarse clustering better than A2's coarse and fine clusterings agree (3.36, not shown in the table). This is promising, but we do not have confidence that the positional dimension is being captured especially well in this model; for those debate-sides labeled *liberal* or *conservative* by A2, the best match of our two positions was still only in agreement only about 60% of the time, and agreement with each human annotator is within the interval of what would be expected if each debate's sides were assigned uniformly at random to positions.[15]

Remarks. Within debates and within topics, the model uses the position variable to distinguish sides well. For external text, the model performs well on articles such as blogs and editorials but on others the positional categories do not seem meaning-

American politics.

[14] In a small number of cases, an annotator declined to label a side. Each unlabeled item received its own cluster.

[15] This was determined using a Monte Carlo simulation with 1,000 samples.

Topic	$i = 1$	$i = 2$
None (ϕ^{I})	vice president, c sections, twenty four, cross pressures, pre dates, anti ballistic, cost effectiveness, anti landmine, court appointed, child poverty	cross examination, under runs, hand outs, half million, non christians, break down, counter argument, seventy five, co workers, run up
"Israel-Palestine"	pre emptive, israeli palestinian, open and shut, first time, hamas controlled, democratically elected	two state, long term, self destructive, secretary general, right wing, all out, near daily, short term
"Same-sex marriage"	same sex, long term, second class, blankenhorn rauch, wrong headed, self denial, left handed	opposite sex, well intentioned, day time, planet wide, day night, child rearing, low earth, one way, one third
"Drugs"	hands free, performance enhancing, in depth, hand held, best kept, non pharmaceutical, anti marijuana	long term, high speed, short term, peer reviewed, alcohol related, mind altering, inner city, long lasting
"Healthcare"	single payer, so called, self sustaining, public private, for profit, long run, high cost, multi payer	government run, government approved, high risk, two tier, government appointed, low cost, set up
"Death penalty"	anti death, non violent, african american, self help, cut and cover, heavy handed, dp equivalent	semi automatic, high profile, hate crime, assault weapons, military style, high dollar, self protective
"Abortion"	pro choice, pro life, non muslim, well educated, anti abortion, much needed, church state, birth control	would be, full time, late term, judeo christian, life style, day to day, non christian, child bearing

Table 7: General position (first row) and topic-specific position bigrams associated with six selected topics.

Topic	Terms	Person entity mentions
"Israel-Palestine"	israel, gaza, hamas, israeli, palestinian	Benjamin Netanyahu, Al Jazeera, Mavi Marmara, Nicholas Kristoff, Steven R. David
"Same-sex marriage"	marriage, gay, mars, space, moon	Buzz Aldrin, Andrew Sullivan, Moon Base, Scott Bidstrup, Ted Olson
"Drugs"	marijuana, drug, drugs, alcohol, age	Four Loko, Evo Morales, Toni Meyer, Sean Flynn, Robert Hahn
"Healthcare"	health, care, insurance, public, private	Kent Conrad, Paul Hsieh, Paul Krugman, Ezra Klein, Jacob Hacker
"Death penalty"	death, crime, punishment, penalty, justice	Adam Bedau, Thomas R. Eddlem, Jeff Jacoby, John Baer, Peter Bronson
"Abortion"	women, religious, abortion, god, life	Ronald Reagan, John Paul II, Sara Malkani, Mother Teresa, Marcella Alsan

Table 8: For 6 selected topics (labels assigned manually), top terms (ϕ^{t}) and person entities (ϕ^{e}). Bigrams were included but did not rank in the top five for these topics. The model has conflated debates relating to same-sex marriage with the space program.

ful, perhaps due to the less argumentative nature of other kinds of articles. Noting the vast literature focusing on ideological positions expressed in text, we believe this failure suggests (i) that broad-based positions that hold across many topics may require richer textual representations (see, e.g., the "syntactic priming" of Greene and Resnik, 2009), or (ii) that an alternative representation of positions, such as the spatial models favored by political scientists (Poole and Rosenthal, 1991), may be more discoverable. Aside from those issues, a stronger theory of positions may be required. Such a theory could be encoded in a more informative prior or weaker independence assumptions across debates. Finally, exploiting explicitly ideological texts alongside the moderated arguments of Debatepedia might also help to identify textual associations with general positions (Sim et al., 2013). We leave these directions to future work.

4.2 Qualitative Analysis

Of the $T = 40$ topics our model inferred, we subjectively judged 37 to be coherent; a glimpse of each is given in Figure 5. We manually selected six of the most interpretable topics for further evaluation.

As a generative modeling approach, our model was designed for the purpose of reducing the dimensionality of the sociopolitical debate space, as evidenced by Debatepedia. It is like other topic models in this regard, but we believe that some effects of our design choices are noteworthy. Table 6 compares the positional bigrams of our model to the sentiments inferred by JST. We observe the benefit of our model in identifying terms associated with positions on social issues, while JST selects more general sentiment terms.

Table 7 shows bigrams most strongly associated with general position distributions ϕ^{i} and selected topic-position distributions ϕ°.[16] We see the potential benefit of multiword expressions. Although we have used frequent bigrams as a poor man's approximation to multiword expression analysis, we find the topic-specific positions terms to be subjectively evocative. While somewhat internally coherent, we do not observe consistent alignment across topics, and the general distributions ϕ^{i} are not suggestive.

The separation of personal name mentions into their own distributions, shown for some topics in Table 8, gives a distinctive characterization of topics based on relevant personalities. Subjectively, the top individuals are relevant to the subject matter associated with each topic (though the topics are not always pure; same-sex marriage and the space program are merged, for example).

5 Related Work

Insofar as debates are subjective, our study is related to **opinion mining**. Subjective text classification (Wiebe and Riloff, 2005) leads to opinion mining tasks such as opinion extraction (Dave et al., 2003), positive and negative polarity classification (Pang et al., 2002), sentiment target detection (Hu and Liu, 2004; Ganapathibhotla and Liu, 2008), and feature-opinion extraction (Wu et al., 2009). The above studies are conducted mostly on product reviews, a domain with a simpler opinion landscape and more concrete rationales for those opinions, compared to sociopolitical debates.

Generative **topic models** have been successfully implemented in opinion mining tasks such as feature identification (Titov and McDonald, 2008), entity-topic extraction (Newman et al., 2006), mining contentious expressions and interactions (Mukherjee and Liu, 2012) and specific aspect-opinion word extraction from labeled data (Zhao et al., 2010). Most relevant to this research is work on feature-sentiment extraction (Lin and He, 2009; Mei et al., 2007). Mei et al. (2007) built on PLSI, which is problematic for generalizing beyond the training sample. The JST model of Lin and He (2009) is an LDA-based topic model in which each word token is assigned both a sentiment and a topic; they exploited a sentiment lexicon in the prior distribution. Our model is closely related, but introduces a switching variable that assigns *some* tokens to positions, some to topics, and some to both. Unlike Lin and He's sentiments, our model's positions are associated with the two sides of a debate, and we incorporate topics at the level of questions within debates.

Some studies have specifically analyzed **contrastive viewpoints** or **stances** in general discussion text. newciteAgrawal03miningnewsgroups used graph mining based method to classify authors in to opposite camps for a given topic. Paul et al. (2010) developed an unsupervised method for summarizing contrastive opinions from customer reviews. Abu-Jbara et al. (2012) and Dasigi et al. (2012) developed techniques to address the problem of automatically detecting subgroups of people holding similar stances in a discussion thread.

Several prior studies have considered **debates**. Cabrio and Villata (2012) developed a system based on argumentation theory which recognizes the entailment and contradiction relationships between two texts. Awadallah et al. (2011) used a debate corpus as a seed for extracting person-opinion-topic tuples from news and other web documents and in later work classified the quotations to specific topics and polarity using language models (Awadallah et al., 2012). Somasundaran and Wiebe (2009) and Anand et al. (2011) were interested in ideological content in debates, relying on discourse structure and leveraging sentiment lexicons to recognize stances.

Closer to the methodology we describe, Lin et al. (2008) presented a statistical model for political discourse that incorporates both topics and ideologies; they used debates on the Israeli-Palestinian conflict. Fortuna et al. (2009) showed that it is possible to isolate a subset of terms from media content that are informative of a news organization's bias towards a particular issue. Ahmed and Xing (2010) introduced multi-level latent Dirichlet allocation, and Eisenstein et al. (2011) introduced sparse additive generative models, both conceived as extensions to well-established probabilistic modeling techniques (Blei et al., 2003); these were applied to debates and political blog datasets. Our approach builds on these models (especially the switching variables of Ahmed and Xing). We go farther in jointly modeling

[16] For more topics, please refer to the supplementary notes.

text across *many* debates evidenced by the structure of Debatepedia, thus grounding our models more solidly in familiar sociopolitical issues, and in making extensive use of existing NLP resources.

6 Conclusion

Using text from Debatepedia, we inferred topics and position term lexicons in the domain of sociopolitical debates. Our approach brings together tools from information extraction and sentiment analysis into a latent-variable topic model and exploits the hierarchical structure of the dataset. Our qualitative and quantitative evaluations show the model's strengths and weaknesses.

Acknowledgments

The authors thank several anonymous reviewers, Justin Gross, David Kaufer, and members of the ARK group at CMU for helpful feedback on this work and gratefully acknowledge the assistance of the annotators. This research is supported by the Singapore National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme Office, by an A*STAR fellowship to Y.S., and by Google's support of the Reading is Believing project at CMU.

References

Amjad Abu-Jbara, Mona Diab, Pradeep Dasigi, and Dragomir Radev. 2012. Subgroup detection in ideological discussions. In *Proceedings of ACL*.

Amr Ahmed and Eric P. Xing. 2010. Staying informed: supervised and semi-supervised multi-view topical analysis of ideological perspective. In *Proceedings of EMNLP*.

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: classifying stance in online debate. In *Proceedings of the Second Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*.

Rawia Awadallah, Maya Ramanath, and Gerhard Weikum. 2011. OpinioNetIt: Understanding the opinions-people network for politically controversial topics. In *Proceedings of CIKM*.

Rawia Awadallah, Maya Ramanath, and Gerhard Weikum. 2012. PolariCQ: Polarity classification of political quotations. In *Proceedings of CIKM*.

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

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of ACL*.

Sarah Cohen, James T. Hamilton, and Fred Turner. 2011. Computational journalism. *Communications of the ACM*, 54(10):66–71.

Pradeep Dasigi, Weiwei Guo, and Mona Diab. 2012. Genre independent subgroup detection in online discussion threads: a pilot study of implicit attitude using latent textual semantics. In *Proceedings of ACL*.

Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of WWW*.

Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text. In *Proceedings of ICML*.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL*.

Blaz Fortuna, Carolina Galleguillos, and Nello Cristianini. 2009. Detecting the bias in media with statistical learning methods. In Ashok N . Srivastava and Mehran Sahami, editors, *Text Mining: Classification, Clustering, and Applications*, pages 27–50. Chapman & Hall/CRC.

Murthy Ganapathibhotla and Bing Liu. 2008. Mining opinions in comparative sentences. In *Proceedings of COLING*.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of HLT-NAACL*.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of CIKM*.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. In *Proceedings of WSDM*.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of CIKM*.

Wei-Hao Lin, Eric Xing, and Alexander Hauptmann. 2008. A joint topic and perspective model for ideological discourse. In *Proceedings of ECML-PKDD*.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of WWW*.

Marina Meila. 2003. Comparing clusterings by the variation of information. In Bernhard Schölkopf and Manfred K. Warmuth, editors, *Learning Theory and Kernel Machines*, volume 2777 of *Lecture Notes in Computer Science*, pages 173–187. Springer.

Arjun Mukherjee and Bing Liu. 2012. Mining contentions from discussions and debates. In *Proceedings of KDD*.

David Newman, Chaitanya Chemudugunta, and Padhraic Smyth. 2006. Statistical entity-topic models. In *Proceedings of KDD*.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of EMNLP*.

Michael J. Paul, ChengXiang Zhai, and Roxana Girju. 2010. Summarizing contrastive viewpoints in opinionated text. In *Proceedings of EMNLP*.

Keith Poole and Howard Rosenthal. 1991. Patterns of congressional voting. *American Journal of Political Science*, pages 118–178.

Yanchuan Sim, Brice Acree, Justin H. Gross, and Noah A. Smith. 2013. Measuring ideological proportions in political speeches. In *Proceedings of EMNLP*.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of ACL*.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of WWW*.

Teun A. Van Dijk. 1998. *Ideology: A Multidisciplinary Approach*. Sage Publications Limited.

Ellen M. Voorhees. 1999. The trec-8 question answering track report. In *Proceedings of TREC*.

Hanna M. Wallach. 2006. Topic modeling: beyond bag-of-words. In *Proceedings of ICML*.

Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of CICLing*.

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Opinionfinder: a system for subjectivity analysis. In *Proceedings of HLT-EMNLP*.

Yuanbin Wu, Qi Zhang, Xuanjing Huang, and Lide Wu. 2009. Phrase dependency parsing for opinion mining. In *Proceedings of EMNLP*.

Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of EMNLP*.