9-2008

A Joint Topic and Perspective Model for Ideological Discourse

Wei-Hao Lin
Carnegie Mellon University

Eric Xing
Carnegie Mellon University

Alexander Hauptmann
Carnegie Mellon University

A Joint Topic and Perspective Model for Ideological Discourse

Wei-Hao Lin, Eric Xing, and Alexander Hauptmann[*]

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213 U.S.A.
{whlin,epxing,alex}@cs.cmu.edu

Abstract. Polarizing discussions on political and social issues are common in mass and user-generated media. However, computer-based understanding of ideological discourse has been considered too difficult to undertake. In this paper we propose a statistical model for ideology discourse. By ideology we mean "a set of general beliefs socially shared by a group of people." For example, Democratic and Republican are two major political ideologies in the United States. The proposed model captures lexical variations due to an ideological text's topic and due to an author or speaker's ideological perspective. To cope with the non-conjugacy of the logistic-normal prior we derive a variational inference algorithm for the model. We evaluate the proposed model on synthetic data as well as a written and a spoken political discourse. Experimental results strongly support that ideological perspectives are reflected in lexical variations.

1 Introduction

When people describe a set of ideas as "ideology", the ideas are usually regarded as false beliefs. Marxists associate the dominant class's viewpoints as ideology. Ideology's pejorative connotation is usually used to describe other group's ideas and rarely our own ideas.

In this paper we take a definition of ideology broader than the classic Marxists' definition, but define ideology as "a set of general beliefs socially shared by a group of people" [1]. Groups whose members share similar goals or face similar problems usually share a set of beliefs that define membership, value judgment, and action. These collective beliefs form an ideology. For example, Democratic and Republican are two major political ideologies in the United States.

Written and spoken discourses are critical in the van Dijk's theory of ideology [1]. Ideology is not innate and must be learned through interaction with

---

the world. Spoken and written texts are major media through which an ideology is understood, transmitted, and reproduced. For example, two presidential candidates, John Kerry and George W. Bush, gave the following answers during a presidential debate in 2004:

Example 1. Kerry: What is an article of faith for me is not something that I can legislate on somebody who doesn't share that article of faith. I believe that choice is a woman's choice. It's between a woman, God and her doctor. And that's why I support that.

Example 2. Bush: I believe the ideal world is one in which every child is protected in law and welcomed to life. I understand there's great differences on this issue of abortion, but I believe reasonable people can come together and put good law in place that will help reduce the number of abortions.

From their answers we can clearly understand their attitude on the abortion issue.

Interest in computer based understanding of ideology dates back to the sixties in the last century, but the idea of learning ideology *automatically* from texts has been considered almost impossible. Abelson expressed a very pessimistic view on automatic learning approaches in 1965 [2]. We share Abelson's vision but do not subscribe to his view. We believe that ideology can be statistically modeled and learned from a large number of ideological texts.

– In this paper we develop a statistical model for ideological discourse. Based on the empirical observation in Section 2 we hypothesize that ideological perspectives were reflected in lexical variations. Some words were used more frequently because they were highly related to an ideological text's topic (i.e., *topical*), while some words were used more frequently because authors holding a particular ideological perspective chose so (i.e., *ideological*).
– We formalize the hypothesis and proposed a statistical model for ideological discourse in Section 3. Lexical variations in ideological discourse were encoded in a word's topical and ideological weights. The coupled weights and the non-conjugacy of the logistic-normal prior posed a challenging inference problem. We develop an approximate inference algorithm based on the variational method in Section 3.2.
Such a model can not only uncover topical and ideological weights from data and can predict the ideological perspective of a document. The proposed model will allow news aggregation service to organize and present news by their ideological perspectives.
– We evaluate the proposed model on synthetic data (Section 4.1) as well as on a written text and a spoken text (Section 4.2). In Section 4.3 we show that the proposed model automatically uncovered many discourse structures in ideological discourse.
– In Section 4.4 we show that the proposed model fit ideological corpora better than a model that assumes no lexical variations due to an author or speaker's ideological perspective. Therefore the experimental results strongly suggested that ideological perspectives were reflected in lexical variations.

2 Motivation

Lexical variations have been identified as a "major means of ideological expression" [1]. In expressing a particular ideological perspective, word choices can highly reveal an author's ideological perspective on an issue. "One man's terrorist is another man's freedom fighter." Labeling a group as "terrorists" strongly reveal an author's value judgement and ideological stance [3].

We illustrate lexical variations in an ideological text about the Israeli-Palestinian conflict (see Section 4.2). There were two groups of authors holding contrasting ideological perspectives (i.e., Israeli vs. Palestinian). We count the words used by each group of authors and showed the top 50 most frequent words in Figure 1.

abu agreement american arab arafat bank bush conflict disengagement fence gaza government international iraq israel israeli israelis israels jerusalem jewish leadership minister palestine palestinian palestinians peace plan political president prime process public return roadmap security settlement settlements sharon sharons solution state states terrorism time united violence war west world years

american arab arafat authority bank conflict elections end gaza government international israel israeli israelis israels jerusalem land law leadership military minister negotiations occupation palestine palestinian palestinians peace people plan political prime process public rights roadmap security settlement settlements sharon side solution state states territories time united violence wall west world

Fig. 1: The top 50 most frequent words used by the Israeli authors (left) and the Palestinian authors (right) in a document collection about the Israeli-Palestinian conflict. A word's size represents its frequency: the larger, the more frequent.

Both sides share many words that are highly related to the corpus's topic (i.e., the Israeli-Palestinian conflict): "Palestinian", "Israeli", "political", "peace", etc. However, each ideological perspective seems to emphasize (i.e., choosing more frequently) different subset of words. The Israeli authors seem to use more "disengagement", "settlement", and "terrorism". On the contrary, the Palestinian authors seem to choose more "occupation", "international", and "land." Some words seem to be chosen because they are about a topic, while some words are chosen because of an author's ideological stance.

We thus hypothesize that lexical variations in ideological discourse are attributed to both an ideological text's topic and an author or speaker's ideological point of view. Word frequency in ideological discourse should be determined by how much a word is related to a text's topic (i.e., *topical*) and how much authors holding a particular ideological perspective emphasize or de-emphasize the word (i.e., *ideological*). A model for ideological discourse should take both topical and ideological aspects into account.

3 A Joint Topic and Perspective Model

We propose a statistical model for ideological discourse. The model associates *topical* and *ideological* weights to each word in the vocabulary. Topical weights represent how frequently a word is chosen because of a text's topic independent of an author or speaker's ideological perspective. Ideological weights, on the other hand, modulate topical weights based on an author or speaker's ideological perspective. To emphasize a word (i.e., choosing the word more frequently) we put a larger ideological weight on the word.
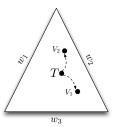

Fig. 2: A three-word simplex illustrates how topical weights T are modulated by two differing ideological weights.

We illustrate the interaction between topical and ideological weights in a three-word simplex in Figure 2. A point T represents topical weights about a specific topic. Suppose authors holding a particular perspective emphasize the word w_3, while authors holding the contrasting perspective emphasize the word w_1. Ideological weights associated with the first perspective will move a multinomial distribution's parameter from T to a new position V_1, which is more likely to generate w_3 than T is. Similarly, ideological weights associated with the second perspective will move the multinomial distribution's parameter from T to V_2, which is more likely to generate w_1 than T is.

3.1 Model Specification

Formally, we combine a word's topical and ideological weights through a logistic function. The complete model specification is listed as follows,

$$P_d \sim \text{Bernoulli}(\pi), d = 1, \ldots, D$$
$$W_{d,n}|P_d = v \sim \text{Multinomial}(\beta_v), n = 1, \ldots, N_d$$
$$\beta_v^w = \frac{\exp(\tau^w \times \phi_v^w)}{\sum_{w'} \exp(\tau^{w'} \times \phi_v^{w'})}, v = 1, \ldots, V$$
$$\tau \sim \text{N}(\mu_\tau, \Sigma_\tau)$$
$$\phi_v \sim \text{N}(\mu_\phi, \Sigma_\phi).$$

We assume that there are two contrasting perspectives in an ideological text (i.e., $V = 2$), and model a document's ideological perspective that its author or speaker holds as a Bernoulli variable P_d, $d = 1, \ldots, D$, where D is the total number of documents in a collection. Each word in a document, $W_{d,n}$, is sampled from a multinomial distribution conditioned on the document d's perspective, $n = 1, \ldots, N_d$, where N_d is a document's length. The bag-of-words representation has been commonly used and shown to be effective in text classification and topic modeling.

The multinomial distribution's parameter, β_v^w, indexed by an ideological perspective v and w-th word in the vocabulary, consists of two parts: topical weights τ and ideological weights ϕ. β is an auxiliary variable, and is deterministically determined by (latent) topical τ and ideological weights $\{\phi_v\}$. The two weights are combined through a logistic function. The relationship between topical and ideological weights is assumed to be multiplicative. Therefore, a word of an ideological weight $\phi = 1$ means that the word is not emphasized or de-emphasized. The prior distributions for topical and ideological weights are normal distributions. The parameters of the joint topic and perspective model, denoted as Θ, include: $\pi, \mu_\tau, \Sigma_\tau, \mu_\phi, \Sigma_\phi$. We call this model a Joint Topic and Perspective Model (jTP). We show the graphical representation of the joint topic and perspective model in Figure 3.
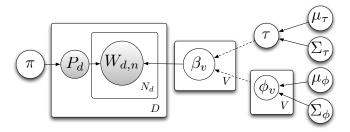

Fig. 3: A joint topic and perspective model in a graphical model representation (see Section 3 for details). A dashed line denotes a deterministic relation between parent and children nodes.

3.2 Variational Inference

The quantities of most interest in the joint topic and perspective model are (unobserved) topical weights τ and ideological weights $\{\phi_v\}$. Given a set of D documents on a particular topic from differing ideological perspectives $\{P_d\}$, the joint posterior probability distribution of the topical and ideological weights

under the joint topic and perspective model is

$$P(\tau, \{\phi_v\}|\{W_{d,n}\}, \{P_d\}; \Theta)$$

$$\propto P(\tau|\mu_\tau, \Sigma_\tau) \prod_v P(\phi_v|\mu_\phi, \Sigma_\phi) \prod_{d=1}^{D} P(P_d|\pi) \prod_{n=1}^{N_d} P(W_{d,n}|P_d, \tau, \{\phi_v\})$$

$$= \mathrm{N}(\tau|\mu_\tau, \Sigma_\tau) \prod_v \mathrm{N}(\phi_v|\mu_\phi, \Sigma_\phi) \prod_d \mathrm{Bernoulli}(P_d|\pi) \prod_n \mathrm{Multinomial}(W_{d,n}|P_d, \beta),$$

where $N(\cdot)$, Bernoulli$(\cdot)$ and Multinomial$(\cdot)$ are the probability density functions of multivariate normal, Bernoulli, and multinomial distributions, respectively.

The joint posterior probability distribution of τ and $\{\phi_v\}$, however, are computationally intractable because of the non-conjugacy of the logistic-normal prior. We thus approximate the posterior probability distribution using a variational method [4], and estimate the parameters using variational expectation maximization [5]. By the Generalized Mean Field Theorem (GMF) [6], we can approximate the joint posterior probability distribution of τ and $\{\phi_v\}$ as the product of individual functions of τ and ϕ_v:

$$P(\tau, \{\phi_v\}|\{P_d\}, \{W_{d,n}\}; \Theta) \approx q_\tau(\tau) \prod_v q_{\phi_v}(\phi_v), \tag{1}$$

where $q_\tau(\tau)$ and $q_{\phi_v}(\phi_v)$ are the posterior probabilities of the topical and ideological weights conditioned on the random variables on their Markov blanket.

Specifically, q_ϕ is defined as follows,

$$q_\tau(\tau) = P(\tau|\{W_{d,n}\}, \{P_d\}, \{\langle\phi_v\rangle\}; \Theta) \tag{2}$$

$$\propto P(\tau|\mu_\tau, \Sigma_\tau) \prod_v P(\langle\phi_v\rangle|\mu_\phi, \Sigma_\phi) P(\{W_{d,n}\}|\tau, \{\langle\phi_v\rangle\}, \{P_d\}) \tag{3}$$

$$\propto \mathrm{N}(\tau|\mu_\tau, \Sigma_\tau) \mathrm{Multinomial}(\{W_{d,n}\}|\{P_d\}, \tau, \{\langle\phi_v\rangle\}), \tag{4}$$

where $\langle\phi_v\rangle$ denotes the GMF message based on $q_{\phi_v}(\cdot)$. From (3) to (4) we drop the terms unrelated to τ.

Calculating the GMF message for τ from (4) is computationally intractable because of the non-conjugacy between multivariate normal and multinomial distributions. We follow the similar approach in [7], and made a Laplace approximation of (4). We first represent the word likelihood $\{W_{d,n}\}$ as the following exponential form:

$$P(\{W_{d,n}\}|\{P_d\}, \tau, \{\langle\phi_v\rangle\}) = \exp\left(\sum_v n_v(\langle\phi_v\rangle \bullet \tau) - \sum_v n_v^T \mathbf{1} C(\langle\phi_v\rangle \bullet \tau)\right) \tag{5}$$

where $\bullet$ is element-wise vector product, n_v is a word count vector under the ideological perspective v, $\mathbf{1}$ is a column vector of one, and C function is defined as follows,

$$C(x) = \log\left(1 + \sum_{p=1}^{P} \exp x_p\right), \tag{6}$$

where P is the dimensionality of the vector x.

We expand C using Taylor series to the second order around $\hat{x}$ as follows,

$$C(x) \approx C(\hat{x}) + \nabla(x)(x - \hat{x}) + \frac{1}{2}(x - \hat{x})^T H(\hat{x})(x - \hat{x}),$$

where ∇ is the gradient of C, and H is the Hessian matrix of C. We set $\hat{x}$ as $\langle \tau \rangle^{(t-1)} \bullet \langle \phi_v \rangle$. The superscript denoted the GMF message in the $t - 1$ (i.e., previous) iteration.

Finally, we plug the second-order Taylor expansion of C back to (4) and rearranged terms about τ. We obtain the multivariate normal approximation of $q_\tau(\cdot)$ with a mean vector μ^* and a variance matrix Σ^* as follows,

$$\Sigma^* = \left(\Sigma_\tau^{-1} + \sum_v n_v^T \mathbf{1} \langle \phi_v \rangle \downarrow H(\hat{\tau} \bullet \langle \phi_v \rangle) \rightarrow \langle \phi_v \rangle \right)^{-1}$$

$$\mu^* = \Sigma^* \left(\Sigma_\tau^{-1} \mu_\tau + \sum_v n_v \bullet \langle \phi_v \rangle - \sum_v n_v^T \mathbf{1} \nabla C(\hat{\tau} \bullet \langle \phi_v \rangle) \bullet \langle \phi_v \rangle \right.$$
$$\left. + \sum_v n_v^T \mathbf{1} \langle \phi_v \rangle \bullet (H(\hat{\tau} \bullet \langle \phi_v \rangle)(\hat{\tau} \bullet \langle \phi_v \rangle)) \right),$$

where $\downarrow$ is column-wise vector-matrix product, $\rightarrow$ is row-wise vector-matrix product. The Laplace approximation for the logistic-normal prior has been shown to be tight [8].

q_{ϕ_v} in (1) can be approximated in a similar fashion as a multivariate normal distribution with a mean vector $\mu^\dagger$ and a variance matrix $\Sigma^\dagger$ as follows,

$$\Sigma^\dagger = \left(\Sigma_\phi^{-1} + n_v^T \mathbf{1} \langle \tau \rangle \downarrow H(\langle \tau \rangle \bullet \hat{\phi}_v) \rightarrow \langle \tau \rangle \right)^{-1}$$

$$\mu^\dagger = \Sigma^\dagger \left(\Sigma_\phi^{-1} \mu_\phi + n_v \bullet \langle \tau \rangle - n_v^T \mathbf{1} \nabla C(\langle \tau \rangle \bullet \hat{\phi}_v) \bullet \langle \tau \rangle \right.$$
$$\left. + n_v^T \mathbf{1} \langle \tau \rangle \bullet (H(\langle \tau \rangle \bullet \hat{\phi}_v)(\langle \tau \rangle \bullet \hat{\phi}_v)) \right),$$

where we set $\hat{\phi}_v$ as $\langle \phi_v \rangle^{(t-1)}$.

In E-step, we have a message passing loop and iterate over the q functions in (1) until converge. We monitor the change in the auxiliary variable β and stop when the absolute change is smaller than a threshold. In M-step, π can be easily maximized by taking the sample mean of $\{P_d\}$. We monitor the data likelihood and stop the variational EM loop when the change of data likelihood is less than a threshold.

3.3 Identifiability

The joint topic and perspective model as specified above is not identifiable. There are multiple assignments of topical and ideological weights that can produce exactly the same data likelihood. Therefore, topic and ideological weights estimated from data may be incomparable.

The first source of un-identifiability is due to the multiplicative relationship between τ and ϕ_v. We can easily multiply a constant to τ^w and divide ϕ_v^w by the same constant, and the auxiliary variable β stays the same.

The second source of un-identifiability comes from the sum-to-one constraint in the multinomial distribution's parameter β. Given a vocabulary $\mathcal{W}$, we have only $|\mathcal{W}| - 1$ number of free parameters for τ and $\{P_d\}$. Allowing $|\mathcal{W}|$ number of free parameters makes topical and ideological weights unidentifiable.

We fix the following parameters to solve the un-identifiability issue: τ^1, $\{\phi_1^w\}$, and ϕ_v^1. We fix the values of the τ^1 to be one and $\{\phi_v^1\}$ to be zero, $v = 1, \ldots, V$. We choose the first ideological perspective as a base and fix its ideological weights ϕ_1^w to be one for all words, $w = 1, \ldots, |\mathcal{W}|$. By fixing the corner of ϕ (i.e., $\{\phi_v^1\}$) we assume that the first word in the vocabulary are not biased by either ideological perspectives, which may not be true. We thus add a dummy word as the first word in the vocabulary, whose frequency is the average word frequency in the whole collection and conveys no ideological information (in the word frequency).

4   Experiments

4.1   Synthetic Data

We first evaluate the proposed model on synthetic data. We fix the values of the topical and ideological weights, and generated synthetic data according to the generative process in Section 3. We test if the variational inference algorithm for the joint topic and perspective model in Section 3.2 successfully converges. More importantly, we test if the variational inference algorithm can correctly recover the true topical and ideological weights that generated the synthetic data.

Specifically, we generate the synthetic data with a three-word vocabulary and topical weights $\tau = (2, 2, 1)$, shown as $\circ$ in the simplex in Figure 4. We then simulate different degrees to which authors holding two contrasting ideological beliefs emphasized words. We let the first perspective emphasize w_2 ($\phi_1 = (1, 1 + p, 0)$) and let the second perspective emphasized w_1 ($\phi_2 = (1 + p, 1, 0)$). w_3 is the dummy word in the vocabulary. We vary the value of p ($p = 0.1, 0.3, 0.5$) and plotted the corresponding auxiliary variable β in the simplex in Figure 4. We generate the equivalent number of documents for each ideological perspective, and varied the number of documents from 10 to 1000.

We evaluate how closely the variational inference algorithm recovered the true topical and ideological weights by measuring the maximal absolute difference between the true β (based on the true topical weights τ and ideological weights $\{\phi_v\}$) and the estimated $\hat{\beta}$ (using the expected topical weights $\langle \tau \rangle$ and ideological weights $\{\langle \phi_v \rangle\}$ returned by the variational inference algorithm).

The simulation results in Figure 5 suggested that the proposed variational inference algorithm for the joint topic and perspective is valid and effective. Although the variational inference algorithm was based on Laplace approximation, the inference algorithm recovered the true weights very closely. The absolute difference between true β and estimated $\hat{\beta}$ was small and close to zero.
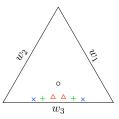
Fig. 4: We generate synthetic data with a three-word vocabulary. The ∘ indicates the value of the true topical weight τ. $\triangle$, $+$, and $\times$ are β after τ is modulated by different ideological weights $\{\phi_v\}$.
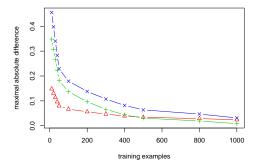

Fig. 5: The experimental results of recovering true topical and ideological weights. The x axis is the number of training examples, and the y axis is the maximal absolute difference between true β and estimated $\hat{\beta}$. The smaller the difference, the better. The curves in $\triangle$, $+$, and $\times$ correspond to the three different ideological weights in Figure 4.

4.2   Ideological Discourse

We evaluate the joint topic and perspective model on two ideological discourses. The first corpus, bitterlemons, is comprised of editorials written by the Israeli and Palestinian authors on the Israeli-Palestinian conflict. The second corpus, presidential debates, is comprised of spoken words from the Democratic and Republican presidential candidates in 2000 and 2004.

The bitterlemons corpus consists of the articles published on the website http://bitterlemons.org/. The website is set up to "contribute to mutual understanding [between Palestinians and Israelis] through the open exchange of ideas."[1] Every week an issue about the Israeli-Palestinian conflict is selected for discussion (e.g., "Disengagement: unilateral or coordinated?"). The website editors have labeled the ideological perspective of each published article. The bitterlemons corpus has been used to learn individual perspectives [9], but the

---

[1] http://www.bitterlemons.org/about/about.html

previous work was based on naive Bayes models and did not simultaneously model topics and perspectives.

The 2000 and 2004 presidential debates corpus consists of the spoken transcripts of six presidential debates and two vice-presidential debates in 2000 and 2004. We downloaded the speech transcripts from the American Presidency Project[2]. The speech transcripts came with speaker tags, and we segmented the transcripts into spoken documents according to speakers. Each spoken document was either an answer to a question or a rebuttal. We discarded the words from moderators, audience, and reporters.

We choose these two corpora for the following reasons. First, the two corpora contain political discourse with strong ideological differences. The bitterlemons corpus contains the Israeli and the Palestinian perspectives; the presidential debates corpus the Republican and Democratic perspectives. Second, they are from multiple authors or speakers. There are more than 200 different authors in the bitterlemons corpus; there are two Republican candidates and four Democratic candidates. We are interested in ideological discourse expressing socially shared beliefs, and less interested in individual authors or candidates' personal beliefs. Third, we select one written text and one spoken text to test how our model behaves on different communication media.

We removed metadata that may reveal an author or speaker's ideological stance but were not actually written or spoken. We removed the publication dates, titles, an author's name and biography in the bitterlemons corpus. We removed speaker tags, debate dates, and location in the presidential debates corpus. Our tokenizer removed contractions, possessives, and cases.

The bitterlemons corpus consists of 594 documents. There are a total of 462308 words, and the vocabulary size is 14197. They are 302 documents written by the Israeli authors and 292 documents written by the Palestinian authors. The presidential debates corpus consists of 1232 spoken documents. There are a total of 122056 words, and the vocabulary size is 16995. There are 235 spoken documents from the Republican candidates, and 214 spoken documents from the Democratic candidates.

4.3 Topical and Ideological Weights

We fit the proposed joint topic and perspective model on two text corpora, and the results were shown in Figure 6 and Figure 7 in color text clouds[3]. Text clouds represent a word's frequency in size. The larger a word's size, the more frequently the word appears in a text collection. Text clouds have been a popular method of summarizing tags and topics on the Internet (e.g., bookmark tags on Del.icio.us[4] and photo tags on Flicker [5]. Here we have matched a word's size with its topical weight τ.

---

[2] http://www.presidency.ucsb.edu/debates.php
[3] We omit the words of low topical and ideological weights due to space limit.
[4] http://del.icio.us/
[5] http://www.flickr.com/

To show a word's ideological weight, we paint a word in color shades. We assign each ideological perspective a color (red or blue). A word's color is determined by which perspective uses a word more frequently than the other. Color shades gradually change from pure colors (strong emphasis) to light gray (no emphasis). The degree of emphasis is measured by how extreme a word's ideological weight ϕ is from one (i.e., no emphasis). Color text clouds allow us to present three kinds of information at the same time: words, their topical weights, and ideological weights.

fence terrorism disengagement terrorist jordan leader case bush jews past appears leaders unilateral jewish forces status iraq arafats line egypt green term arafat level approach abu settlers months left territory good arabs idea large syria suicide war strategic arab back democratic year sharons effect settlements decision bank west agreement majority water present mazen gaza pa sharon minister prime withdrawal israels return state israel process american oslo violence support security ariel peace conflict issue president current israeli sides palestinian israelis solution future middle jerusalem settlement world force plan long make issues time leadership public refugees east political administration pressure palestinians camp strip palestine ceasefire roadmap national policy government final order situation military economic hamas elections part states international end community territories negotiations based agreements real side united recent work 1967 party made movement important control authority dont hand violent borders continue change including clear relations problem society resolution parties building people al means move power role refugee ongoing intifada nations major civilians fact occupation areas talks council land struggle efforts hope position compromise rights stop difficult put historic opinion positions give accept reason inside law internal occupied americans years significant result ending things wall resistance

Fig. 6: Visualize the topical and ideological weights learned by the joint topic and perspective model from the bitterlemons corpus (see Section 4.3). Red: words emphasized more by the Israeli authors. Blue: words emphasized more by the Palestinian authors.

Let us focus on the words of large topical weights learned from the bitterlemons corpus (i.e., words in large sizes in Figure 6). The word of the largest topical weight is "Palestinian", followed by "Israeli", "Palestinians", "peace", and "political". The topical weights learned by the joint topic and perspective model clearly match our expectation from the discussions about the Israeli-Palestinian conflict. Words in large sizes summarizes well what the bitterlemons corpus is about.

Similarly, a brief glance over words of large topical weights learned from the presidential debates corpus (i.e., words in large sizes in Figure 7) clearly tells us the debates' topic. Words of large topical weights capture what American politics is about (e.g., "people", "president", "America", "government") and specific political and social issues (e.g., "Iraq", "taxes", "Medicare"). Although

companies cut john families kids class american governor nuclear give fight gore ago back jim americans history fund oil didnt year country 1 budget cuts job jobs al 000 laden bin agree national lost kerry ill years presidents rights today bush health president parents middle number united choice social children schools left college debt countries day america insurance drug security big bring general things theyve plan school percent weapons program support benefits forces question means care put bill respect states theyre war vice world fact tax thing ive pay problem talk military iraq great trillion im life medicare billion million good public safe congress prescription education time kind people difference terrorists dont wrong long 2 made make hussein change important saddam hes clear drugs senate administration law money working doesnt man spending mr peace making part lead leadership nation high intelligence policy troops government move programs coming destruction child find threat business lot side weve called issue interest youre voted small state seniors energy hard lets afghanistan strong decision qaida thought deal work end local sense set vote marriage terror problems wont protect gun understand federal hope reform system increase nations matter senator talks continue record texas place lives east folks taxes freedom decisions washington citizens free opponent relief youve

Fig. 7: Visualize the topical weights and ideological weights learned by the joint topic and perspective model from the presidential debates corpus i(see Section 4.3). Red: words emphasized by the Democratic candidates. Blue: words emphasized by the Republican candidates.

not every word of large topical weights is attributed to a text's topic, e.g., "im" ("I'm" after contraction is removed) occurred frequently because of the spoken nature of debate speeches, the majority of words of large topical weights appear to convey what the two text collections are about.

Now let us turn our attention to words' ideological weights ϕ, i.e., color shade in Figure 6. The word "terrorism", followed by "terrorist", is painted pure red, which is highly emphasized by the Israeli authors. "Terrorist" is a word that clearly reveals an author's attitude toward the other group's violent behavior. Many words of large ideological weights can be categorized into the ideology discourse structures previously manually identified by researchers in discourse analysis [1]:

- Membership: Who are we and who belongs to us? "Jews" and "Jewish" are used more frequently by the Israeli authors than the Palestinian authors. "Washington" is used more frequently by the Republican candidates than Democratic candidates.
- Activities: What do we do as a group? "Unilateral", "disengagement", and "withdrawal" are used more frequently by the Israeli authors than the Palestinian authors. "Resistance" is used more frequently by the Palestinian authors than the Israeli authors.

- Goals: What is our group's goal? (Stop confiscating) "land" , "independent", and (opposing settlement) "expansion" are used more frequently by the Palestinian authors than the Israeli authors.
- Values: How do we see ourselves? What do we think is important? "Occupation" and (human) "rights" are used more frequently by Palestinian authors than the Israeli authors. "Schools", "environment", and "middle" "class" are used more frequently by the Democratic candidates than the Republican candidates. "Freedom" and "free" are used more frequently by the Republican candidates.
- Position and Relations: what is our position and our relation to other groups? "Jordan" and "Arafats" (after removing contraction of "Arafat's") are used more frequently by the Israeli authors than by the Palestinian authors.

We do not intend to give a detailed analysis of the political discourse in the Israeli-Palestinian conflict and in American politics. We do, however, want to point out that the joint topic and perspective model seems to "discover" words that play important roles in ideological discourse. The results not only support the hypothesis that ideology is greatly reflected in an author or speaker's lexical choices, but also suggest that the joint topic and perspective model closely captures the lexical variations.

Political scientists and media analysts can formulate research questions based on the uncovered topical and ideological weights, such as: what are the important topics in a text collection? What words are emphasized or de-emphasized by which group? How strongly are they emphasized? In what context are they emphasized? The joint topic and perspective model can thus become a valuable tool to explore ideological discourse.

Our results, however, also point out the model's weaknesses. First, a bag-of-words representation is convenient but fails to capture many linguistic phenomena in political discourse. "Relief" is used to represent tax relief, marriage penalty relief, and humanitarian relief. Proper nouns (e.g., "West Bank" in the bitterlemons corpus and "Al Quida" in the presidential debates corpus) are broken into multiple pieces. N-grams do not solve all the problems. The discourse function of the verb "increase" depends much on the context. A presidential candidate can "increase" legitimacy, profit, or defense, and single words cannot distinguish them.

4.4   Prediction

We evaluate how well the joint topic and perspective model predicted words from unseen ideological discourse in terms of perplexity on a held-out set. Perplexity has been a popular metric to assess how well a statistical language model generalizes [10]. A model generalizes well if it achieves lower perplexity. We choose unigram as a baseline. Unigram is a special case of the joint topic and perspective model that assumes *no* lexical variations are due to an author or speaker's ideological perspective (i.e., fixing all $\{\phi_v\}$ to one).

Perplexity is defined as the exponential of the negative log word likelihood with respect to a model normalized by the total number of words:

$$\exp\left(\frac{-\log P(\{W_{d,n}\}|\{P_d\};\Theta)}{\sum_d N_d}\right)$$

We can integrate out topical and ideological weights to calculate the predictive probability $P(\{W_{d,n}\}|\{P_d\};\Theta)$:

$$P(\{W_{d,n}\}|\{P_d\};\Theta) = \int\int\prod_{d=1}^{D}\prod_{n=1}^{N_d} P(W_{d,n}|P_d)d\tau d\phi_v.$$

Instead, we approximate the predictive probability by plugging in the point estimates of τ and ϕ_v from the variational inference algorithm.

For each corpus, we vary the number of training documents from 10% to 90% of the documents, and measured perplexity on the remaining 10% held-out set. The results were shown in Figure 8. We can clearly see that the joint topic and perspective model reduces perplexity on both corpora. The results strongly support the hypothesis that ideological perspectives are reflected in lexical variations. Only when ideology is reflected in lexical variations can we observe the perplexity reduction from the joint topic and perspective model. The results also suggest that the joint topic and perspective model closely captures the lexical variations due to an author or speaker's ideological perspective.
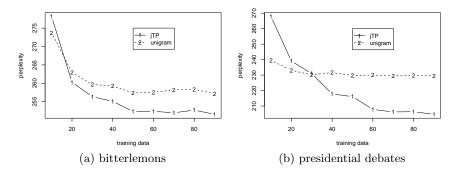

(a) bitterlemons   (b) presidential debates

Fig. 8: The proposed joint topic and perspective model reduces perplexity on a held-out set.

5   Related Work

Abelson and Carroll pioneered modeling ideological beliefs in computers in the sixties [2]. Their system modeled the beliefs of a right-wing politician as a set of

English sentences (e.g., "Cuba subverts Latin America."). Carbonell proposed a system, POLITICS, that can interpret text from two conflicting ideologies [11]. These early studies model ideology at a more sophisticated level (e.g., goals, actors, and action) than the proposed joint topic and perspective model, but require humans to manually construct a knowledge database. The knowledge-intensive approaches suffer from the "knowledge acquisition bottleneck." We take a completely different approach and aim to *automatically* learn ideology from a large number of documents.

[12] explored a similar problem of identifying media's bias. They found that the sources of news articles can be successfully identified based on word choices using Support Vector Machines. They identified the words that can best discriminate two news sources using Canonical Correlation Analysis. In addition to the clearly different methods between [12] and this paper, there are crucial differences. First, instead of applying two different methods as [12] did, the Joint Topic and Perspective Model (Section 3) is a single unified model that can learn to predict an article's ideological slant and uncover discriminating word choices *simultaneously*. Second, the Joint Topic and Perspective Model makes explicit the assumption of the underlying generative process on ideological text. In contrast, discriminative classifiers such as SVM do not model the data generation process [13]. However, our methods implicitly assume that documents are about the same news event or issue, which may not be true and could benefit from an extra story alignment step as [12] did.

We borrow statistically modeling and inference techniques heavily from research on topic modeling (e.g., [14], [15] and [16]). They focus mostly on modeling text collections that containing many *different* (latent) topics (e.g., academic conference papers, news articles, etc). In contrast, we are interested in modeling ideology texts that are mostly on the *same* topic but mainly differs in their ideological perspectives. There have been studies going beyond topics (e.g., modeling authors [17]). We are interested in modeling lexical variation collectively from multiple authors sharing similar beliefs, not lexical variations due to individual authors.

6   Conclusion

We present a statistical model for ideological discourse. We hypothesized that ideological perspectives were partially reflected in an author or speaker's lexical choices. The experimental results showed that the proposed joint topic and perspective model fit the ideological texts better than a model naively assuming no lexical variations due to an author or speaker's ideological perspectives. We showed that the joint topic and perspective model uncovered words that represent an ideological text's topic as well as words that reveal ideological discourse structures. Lexical variations appeared to be a crucial feature that can enable automatic understanding of ideological perspectives from a large amount of documents.

References

1. Van Dijk, T.A.: Ideology: A Multidisciplinary Approach. Sage Publications (1998)
2. Abelson, R.P., Carroll, J.D.: Computer simulation of individual belief systems. The American Behavioral Scientist **8** (May 1965) 24–30
3. Carruthers, S.L.: The Media At War: Communication and Conflict in the Twentieth Century. St. Martin's Press (2000)
4. Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K.: An introduction to variational methods for graphical models. Machine Learning **37** (1999) 183–233
5. Attias, H.: A variational bayesian framework for graphical models. In: Advances in Neural Information Processing Systems 12. (2000)
6. Xing, E.P., Jordan, M.I., Russell, S.: A generalized mean field algorithm for variational inference in exponential families. In: Proceedings of the 19th Annual Conference on Uncertainty in AI. (2003)
7. Xing, E.P.: On topic evolution. Technical Report CMU-CALD-05-115, Center for Automated Learning & Discovery, Pittsburgh, PA (December 2005)
8. Ahmed, A., Xing, E.P.: On tight approximate inference of the logistic-normal topic admixture model. In: Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics. (2007)
9. Lin, W.H., Wilson, T., Wiebe, J., Hauptmann, A.: Which side are you on? identifying perspectives at the document and sentence levels. In: Proceedings of Tenth Conference on Natural Language Learning (CoNLL). (2006)
10. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. The MIT Press (1999)
11. Carbonell, J.G.: POLITICS: Automated ideological reasoning. Cognitive Science **2**(1) (1978) 27–51
12. Fortuna, B., Galleguillos, C., Cristianini, N.: Detecting the bias in media with statistical learning methods. In: Text Mining: Theory and Applications. Taylor and Francis Publisher (2008)
13. Rubinstein, Y.D.: Discriminative vs Informative Learning. PhD thesis, Department of Statistics, Stanford University (January 1998)
14. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (1999) 50–57
15. Blei, D.M., Ng, A.Y., Jordan, M.: Latent dirichlet allocation. Journal of Machine Learning Research **3** (January 2003) 993–1022
16. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences **101** (2004) 5228–5235
17. Rosen-Zvi, M., Griffths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceedings of the 20th Conference on Unvertainty in Artificial Intelligence. (2004)