Detecting Perspectives in Political Debates

David Vilares
Universidade da Coruña
Departamento de Computación
Campus de Elviña s/n, 15071
A Coruña, Spain
david.vilares@udc.es

Yulan He
School of Engineering and Applied Science
Aston University
United Kingdom
y.he@cantab.net

Abstract

We explore how to detect people's perspectives that occupy a certain proposition. We propose a Bayesian modelling approach where topics (or propositions) and their associated perspectives (or viewpoints) are modeled as latent variables. Words associated with topics or perspectives follow different generative routes. Based on the extracted perspectives, we can extract the top associated sentences from text to generate a succinct summary which allows a quick glimpse of the main viewpoints in a document. The model is evaluated on debates from the House of Commons of the UK Parliament, revealing perspectives from the debates without the use of labelled data and obtaining better results than previous related solutions under a variety of evaluations.

1 Introduction

Stance classification is binary classification to detect whether people is supporting or against a topic. Existing approaches largely rely on labelled data collected under specific topics for learning supervised classifiers for stance classification (Mohammad et al., 2016a). At most time, apart from detecting one's stance, we are interested in finding out the arguments behind the person's position. Perspectives, that state people's ideas or the facts known to one, can be *contrastive*, i.e. to be in favour of or against something (e.g. Brexit vs Bremain), or *non-contrastive*, i.e. independent discussions that share a common topic (e.g. unemployment and migration in the context of economy).

Recent years have seen increasing interests in argumentation mining which involves the automatic identification of argumentative structures, e.g., the claims and premises, and detection of argumentative relations between claims and premises or evidences. However, learning models for argumentation mining often require text labelled with components within argumentative structures and detailed indication of argumentative relations among them. Such labelled data is expensive to obtain in practice and it is also difficult to port models trained on one domain to another.

We are particularly interested in detecting different perspectives in political debates. Essentially, we would like to achieve somewhere in between stance classification and argumentation mining. Given a text document, we want to identify a speaker's key arguments, without the use of any labelled data. For example, in debates about '*Education*', we want to automatically extract sentences summarising the key perspectives and their arguments, e.g. 'our education system needs to promote excellence in *stem subjects*', 'teenagers need to be taught with *sexual and health education*' or '*grammar schools* promote inequality'. Similarly, if '*Brexit*' is being discussed in terms of leaving or remaining, we want to cluster arguments into those two viewpoints.

To do this, we introduce a Latent Argument Model (LAM) which assumes that words can be separated as topic words and argument words and follow different generative routes. While topic words only involve a sampling of topics, argument words involve a joint sampling of both topics and arguments. The model does not rely on labelled data as opposed to most existing approaches to stance classification or argument recognition. It is also different from cross-perspective topic models which assume the perspectives are observed (Fang et al., 2012). Quantitative and qualitative evaluations on debates from the House of Commons of United Kingdom show the utility of the approach and provide a comparison against related models.

2 Related work

Our research is related to stance classification, argument recognition and topic modelling for sentiment/perspective detection.

2.1 Stance Classification

Stance detection aims to automatically detect from text whether the author is in favour of, against, or neutral towards a target. As previously reported in (Mohammad et al., 2016b), a person may express the same stance towards a target by using negative or positive language. Hence, stance detection is different from sentiment classification and sentiment features alone are not sufficient for stance detection. With the introduction of the shared task of stance detection in tweets in SemEval 2016 (Mohammad et al., 2016a), there have been increasing interests of developing various approaches for stance detection. But most of them focused on building supervised classifiers from labelled data. The best performing system (Zarrella and Marsh, 2016) made use of large unlabelled data by first learning sentence representations via a hashtag prediction auxiliary task and then fine tuning these sentence representations for stance detection on several hundred labelled examples. Nevertheless, labelled data are expensive to obtain and there is a lack of portability of classifiers trained on one domain to move to another domain.

2.2 Argument Recognition

Closely related to stance detection is argument recognition which can be considered as a more fine-grained task that it aims to identify text segments that contain premises that are against or in support of a claim. Cabrio and Villata (2012) combined textual entailment with argumentation theory to automatically extract the arguments from online debates. Boltuzic and Šnajder (2014) trained supervised classifiers for argument extraction from their manually annotated corpus by collecting comments from online discussions about two specific topics. Sardianos et al. (2015) proposed a supervised approach based on Conditional Random Fields for argument extraction from Greek news. Nguyen and Litman (2015) run an LDA model and post-processed the output, computing argument and domain weights for each of the topics, which were then used to extract argument and domain words. Their model outperformed traditional n-grams and lexical/syntactic rules on a collection of persuasive essays. Lippi and Torroni (2016a) hypothesized that vocal features of speech can improve argument mining and proposed to train supervised classifiers by combining features from both text and speech for claim detection from annotated political debates. Apart from claim/evidence detection, there has also been work focusing on identification of argument discourse structures such as the prediction of relations among arguments or argument components (Stab and Gurevych, 2014; Peldszus and Stede, 2015). A more recent survey of various machine learning approaches used for argumentation mining can be found in (Lippi and Torroni, 2016b). All these approaches have been largely domain-specific and rely on a small set of labelled data for supervised model learning.

2.3 Topic Modeling for Sentiment/Perspective Detection

Topic models can be modified to detect sentiments or perspectives. Lin and He (2009) introduced a joint sentiment topic (JST) model, which simultaneously extracts topics and topic-associated sentiments from text. Trabelsi and Zaıane (2014) proposed a joint topic viewpoint (JTV) model for the detection of latent viewpoints under a certain topic. This is essentially equivalent to the reparameterized version of the JST model called REVERSE-JST (Lin et al., 2012) in which sentiment label (or viewpoint) generation is dependent on topics, as opposed to JST where topic generation is conditioned on sentiment labels.

Fang et al. (2012) proposed a Cross-Perspective Topic Model (CPT) in which the generative processes for topic words (nouns) and opinion words (adjectives, adverbs and verbs) are different, as the opinion words are sampled independently from the topic. Also, CPT assumed perspectives are observed, which implies texts need to be annotated with the viewpoint they belong to. Awadallah et al. (2012) detected politically controversial topics by creating an opinion-base of opinion holders and their views. Das and Lavoie (2014) observed the editions and interactions of a user in Wikipedia pages to infer topics and points of view at the same time. Qiu et al. (2015) proposed a regression-based latent factor model which jointly models user arguments, interactions, and attributes for user stance prediction in online debates.

3 Latent Argument Model (LAM)

We assume that in a political debate, the speaker first decides on which topic she wants to comment on (e.g. *Education*). She then takes a stance (e.g. *remark the importance about stem subjects*) and elaborates her stance with arguments. It is worth noting that we do not consider the temporal dimension of documents here, i.e., our model is fed with a collection of unlabeled documents without temporal order.

We use a switch variable x to denote whether a word is a *background word* (shared across multiple topics), a *topic word* (relating to a certain topic) or an *argument word* (expressing arguments under a specific topic). Depending on the type of word, we follow a different generative process. For each word in a document, if it is a background word, we simply sample it from the background word distribution ϕ^b; if it is a topic word, we first sample a topic z from the document-specific topic distribution θ_d and then sample the word from the topic-word multinomial distribution ψ_z shared across all documents; if it is an argument word, we need to first jointly sample the topic-argument pair, (z, a), where z comes from the existing topics already sampled for the topic words in the document and a is sampled from the topic-specific argument distribution ω_z, and finally the word is sampled from the multinomial word distribution for the topic-specific argument $\psi_{z,a}$. The argument indicator here is a latent categorical variable. It can take a binary value to denote pro/con or positive/negative towards a certain topic. More generally, it could also take a value from multiple stance or perspective categories. We thus propose a Latent Argument Model (LAM) shown in Figure 1. Formally, the generative process is as follows:

- Draw a distribution over the word switch variable, $\phi \sim$ Dirichlet(γ), and background word distribution, $\psi^b \sim$ Dirichlet(β^b).
- For each topic $z \in \{1...T\}$, draw a multinomial topic-word distribution $\psi_z \sim$ Dirichlet(β^z).
  - For each argument $a \in \{1...A\}$ draw a multinomial topic-argument distribution $\omega_z \sim$ Dirichlet(δ) as well as a multinomial topic-argument-word distribution $\psi^v_{z,a} \sim$ Dirichlet(β^a).
- For each document $d \in \{1...D\}$:
  - Draw a multinomial topic distribution, $\theta_d \sim$ Dirichlet(α).
  - For each word $n \in \{1, .., N_d\}$ in d:
    * Choose $x_{d,n} \sim$ Multinomial(ϕ).
    * If $x_{d,n} = 0$, draw a background word $w_{d,n} \sim \psi^b$;
    * If $x_{d,n} = 1$, draw a topic $z \sim$ Multinomial(θ_d) and a word $w_{d,n} \sim$ Multinomial(ψ_z);
    * If $x_{d,n} = 2$, draw a topic $z \sim$ Multinomial(θ_d), an argument $a \sim$ Multinomial(ω_z) and a word $w_{d,n} \sim$ Multinomial$(\psi^a_{z,a})$.

Figure 1 shows its plate representation.

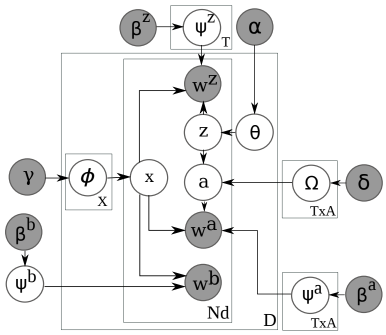

Figure 1: The plate notation for the LAM model. Shadowed elements represent the observed variables (words and prior distributions).

3.1 Inference and Parameter Estimation

We use Collapsed Gibbs Sampling (Casella and George, 1992) to infer the model parameters and the latent assignments of topics and arguments, given the observed data. Gibbs sampling is a Markov chain Monte Carlo method to iterative estimate latent parameters. In each iteration, a new sample of the hidden parameters is made based on the distribution of the previous epoch. Letting the index $t = (d, n)$ denote the nth word in document d and the subscript $-t$ denote a quantity that excludes data from the nth word position in document d, $\Lambda = \{\alpha, \beta^b, \beta^z, \beta^a, \gamma, \delta\}$, the conditional posterior for x_t is:

$$P(x_t = r|\boldsymbol{x}_{-t}, \boldsymbol{z}, \boldsymbol{a}, \boldsymbol{w}, \Lambda) \propto \frac{\{N^r_d\}_{-t} + \gamma}{\{N_d\}_{-t} + 3\gamma} \cdot \frac{\{N^r_{w_t}\}_{-t} + \beta^r}{\sum_{w'}\{N^r_{w'}\}_{-t} + W\beta^r}, \quad (1)$$

where r denotes different word types, either background word, topic word or argument word. N^r_d denotes the number of words in document d assigned to the word type r, N_d is the total number of words in the document d, $N^r_{w_t}$ is the number of times word w_t is sampled from the distribution for the word type r, W is the vocabulary size.

For topic words, the conditional posterior for z_t is:

$$P(z_t = k|\boldsymbol{z}_{-t}, \boldsymbol{w}, \Lambda) \propto \frac{N_{d,k}^{-t} + \alpha_k}{N_d^{-t} + \sum_k \alpha_k} \cdot \frac{N_{k,w_t}^{-t} + \beta^z}{N_k^{-t} + W\beta^z}, \quad (2)$$

where $N_{d,k}$ is the number of times topic k was assigned to some word tokens in document d, N_d is the total number of words in document d, N_{k,w_t} is the number of times word w_t appeared in topic k.

For argument words, the conditional posterior for z_t and a_t is:[1]

$$P(z_t = k, a_t = j|\boldsymbol{z}_{-t}, \boldsymbol{a}_{-t}, \boldsymbol{w}, \Lambda) \propto \frac{N_{d,k}^{-t} + \alpha_k}{N_d^{-t} + \sum_k \alpha_k} \cdot \frac{N_{k,j}^{-t} + \delta_{k,j}}{N_k^{-t} + \sum_j \delta_{k,j}} \cdot \frac{N_{k,j,w_t}^{-t} + \beta^v}{N_{k,j}^{-t} + W\beta^v}, \quad (3)$$

where $N_{k,j}$ is the number of times a word has been associated with the topic k and argument j, N_{k,j,w_t} is the number of times word w_t appeared in topic k and with argument j, and $N_{k,j}$ is the number of words assigned to topic k and argument j.

Once the assignments for all the latent variables are known, we can easily estimate the model parameters $\{\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{\psi^b}, \boldsymbol{\psi^z}, \boldsymbol{\psi^a}, \boldsymbol{\omega}\}$. We set the symmetric prior $\gamma = 0.3$, $\epsilon = 0.01$, $\beta^b = \beta^z = 0.01$, $\delta = (0.05 \times L)/A$, where L is the average document length, A the is total number of arguments, and the value of 0.05 on average allocates 5% of probability mass for mixing. The asymmetric prior $\boldsymbol{\alpha}$ is learned directly from data using maximum-likelihood estimation (Minka, 2003) and updated every 40 iterations during the Gibbs sampling procedure. In this paper we only consider two possible stances, hence, $A = 2$. But the model can be easily extended to accommodate more than two stances or perspectives. We set the asymmetric prior β^a for the topic-argument-word distribution based on a subjectivity lexicon in hoping that contrastive perspectives can be identified based on the use of positive and negative words. We run Gibbs sampler for 1 000 iterations and stop the iteration once the log-likelihood of the training data converges.

[1] This equation had a typo in the original submitted paper that has been corrected.

3.2 Separating Topic and Perspective Words

Using the word type switch variable $\mathbf{x}$, we could separate topic and argument words in LAM based solely on the statistics gathered from data. We also explore another two methods to separate topic and argument words based on Part-of-Speech (POS) tags and with the incorporation of a subjectivity lexicon. For the first variant, we adopt the similary strategy as in (Fang et al., 2012) that nouns (NOUN) are topic words; adjectives (ADJ), adverbs (ADV) and verbs (VERB) are argument words; words with other POS tags are background words. Essentially, $\mathbf{x}$ is observed. We call this model LAM_POS.

For the second variant, instead of assuming $\mathbf{x}$ is observed, we incorporate the POS tags as prior information to modify the Dirichlet prior γ for the word type switch variable at the initialization step. In addition, we also consider a subjective lexicon[2], L, that if a word can be found in the lexicon, then it is very likely the word is used to convey an opinion or argument, although there is still a small probability that word could be either background or topic word. Assuming an asymmetric Dirichlet prior for $\mathbf{x}$ is parametrized by $\gamma^\intercal = [\gamma^b, \gamma^z, \gamma^a]$ for background, topic and argument words, it is modified by a transformation matrix λ, $\gamma^{\text{new}} = \lambda \times \gamma^\intercal$, where λ is defined by:

- If word $w \in L \wedge \text{POSTAG}(w) \neq \text{NOUN}$ then $\lambda^\intercal = [0.05, 0.05, 0.9]$
- else if $\text{POSTAG}(w) = \text{NOUN}$ then $\lambda^\intercal = [0.05, 0.9, 0.05]$
- else if $\text{POSTAG}(w) \in \{\text{ADJ,ADV,VERB}\}$ then $\lambda^\intercal = [0.05, 0.05, 0.9]$
- else $\lambda^\intercal = [0.9, 0.05, 0.05]$

The conditional probability for the switch variable x is modified by simultaneously considering the POS tag g for the word at position t:

$$P(x_t = r, y_t = g|\boldsymbol{x}_{-t}, \boldsymbol{z}, \boldsymbol{a}, \boldsymbol{w}, \Lambda) \propto \frac{\{N_d^r\}_{-t} + \gamma}{\{N_d\}_{-t} + 3\gamma} \cdot \frac{\{N_{w_t}^r\}_{-t} + \beta^r}{\sum_{w'}\{N_{w'}^r\}_{-t} + W\beta^r} \cdot \frac{\{N_g^r\} + \epsilon_g^r}{\{N_g\} + \sum_{g'} \epsilon_{g'}^r}, \quad (4)$$

where an additional term is added to the RHS of Equation 1. Here, N_g^r denotes the number of words with POS tag g assigned to the word type

[2] In this work, we use the subjectivity lexicon presented at (Wiebe et al., 2005).

r, N_g is the total number of words assigned to the POS tag g, ϵ_g^r is the Dirichlet prior for the POS tag-word type distribution.

We call the second variant LAM_LEX. As both the POS tag information and the subjectivity lexicon are only incorporated in the initialisation step, LAM_LEX sits in-between LAM and LAM_POS that it performs soft clustering of topic words and argument words. That is, during the initialisation, nouns are more likely to be topic words, but there is still a small probability that they could be either argument or background words; and similarly for words tagged as adjectives, adverts and verbs.

4 House of Common Debates (HCD)

Debates from the UK parliament are archived and available for consulting.[3] A custom web-crawler was developed to obtain the records of every day that The House of Commons was in session between 2009 and 2016. Due to inconsistencies in the data format and volume of data, much of the analysis focuses on the recordings for the parliamentary year 2014-2015. The general structure of a single day of recording is as follows: a question will be put to the house (generally a Bill) and Members of Parliament (MPs) will discuss various aspects regarding the Bill or show stances about it. Each speech made by an MP is considered to be a single document. Multiple Bills will be discussed each day. The current item being discussed is clearly marked in the source data format, so linking documents to the current bill and MP is trivial. Each speech is labelled with a major (e.g. *education*) and a minor topic (e.g. *grammar schools*) and help us create a dataset with the desired needs.

The length that The House will be in session varies and the number of bills discussed also varies. In this paper, we considered debates occurred during March of 2015[4] and contains 1 992 speeches belonging to diverse domains: *justice*, *education*, *energy and climate change*, *treasury*, *transport*, *armed forces*, *foreign and commonwealth office*, *environment*, *transport*, *royal assent*, *work and pensions*, *northern Ireland*. This House of Commons Debates (HCD) dataset is made available for the research community.[5]

We followed a standard methodology to clean the texts: stopwords were removed, lemmatization was applied, and a naive negation treatment was considered for the particle 'not', by creating bigrams for words occurs in the subjectivity lexicon (e.g., 'not good' becomes 'not_good'). As topic models suffer from lack of robustness if large outliers are present, we also removed very frequent (above 99%) and rare words (below percentile 65%), assuming that word occurrences of the collection follow a Zip's law distribution.[6] Similar strategy was carried out for texts, in order to just consider texts of similar length. The preprocessed HCD contains a total of 1 598 speeches.

[3] https://hansard.parliament.uk

[4] Period of time what selected on a basis of existence of a large number of major topics.

[5] https://github.com/aghie/lam/blob/master/hcd.tsv

5 Experiments

This section evaluates LAM and its variants qualitatively and quantitatively (averaged over 5 runs).

The models for comparison are listed below:

- **LDA**. Latent Dirichlet Allocation (Blei et al., 2003).
- **CPT**. The Cross-perspective Topic Model (Fang et al., 2012) assumes perspectives are observed. To be able to run this model on the political speeches, we implemented a version that can manage latent perspectives and separately sample topics and viewpoints.
- **JTV**. Joint Topic-Viewpoint Model (Trabelsi and Zaıane, 2014) is essentially the reparameterized version of the Joint Sentiment-Topic (JST) model (Lin and He, 2009) called REVERSE-JST (Lin et al., 2012) in which sentiment label (or viewpoint) generation is dependent on topics. We implemented JTV as the reversed JST model.[7]
- **LAM**. Latent Argument Model from §3.
- **LAM_POS**. LAM with topic, argument or background words separated by POS tags.
- **LAM_LEX**. Both POS tags and a subjective lexicon are used to initialise the Dirichlet prior γ for the word type switch variable as described in §3.2.

5.1 Experimental Results

Results are evaluated in terms of both topic coherence and the quality of the extracted perspectives.

5.1.1 Topic Coherence

The CV metric[8] is used to measure the coherence of the topics generated by the models as it has been

[6] Percentiles were selected on an empirical basis.

[7] We were not able to find a publicly available code of the JTV implementation.

[8] https://github.com/AKSW/Palmetto/

shown to give the results closest to human evaluation compared to other topic coherence metrics (Röder et al., 2015). In brief, given a set of words, it gives an intuition of how likely those words co-occur compared to expected by chance.

Figure 2 plots the CV results[9] versus the number of topics on HCD for various models. For each topic z, we extract the top ten most representative words ranked by their respective normalised discriminative score defined by $\mathrm{DS}(w,z) = P(w|z)/[\max_{z' \neq z} P(w|z')]$. We chose this approach instead of simple $P(w|z)$ as it was observed to turn into higher quality topics. It is clear that LAM_LEX models outperform baselines and that all variants are learning well the topics from the data, showing that the three different mechanisms for the switch variable are effective to generate coherent topics. Also, our models work robustly under different number of topics. Moreover, LAM_LEX achieve better coherence scores than the original LAM and LAM_POS. This shows that it is more effective to use POS tags and a subjectivity lexicon to initialise the Dirichlet prior for the word type switch variable rather than simply relying on POS tags or subjectivity lexica to give hard discrimination between topic and argument words.

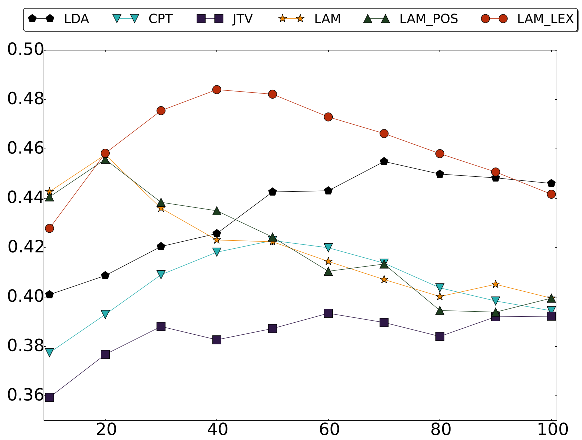

Figure 2: CV coherence vs the number of topics for different modeling approaches.

We also used the gold-standard major topic label assigned by Hansard to each speech to carry out an additional quantitative evaluation. For each topic z, we extract the top ten most representative sentences ranked by their respective normalised discriminative score defined by $\mathrm{DS}(s,z) = \sum_{w \in \boldsymbol{s}} \mathrm{DS}(w,z)/\mathrm{Length}(\boldsymbol{s})$.

If a particular model is clustering robustly, the top sentences it extracts should belong to speeches that discuss the same topic and share the same major and/or minor topic labels in the HCD corpus. Table 2 shows for the studied models the percentage of sentences where x out of top 10 topic sentences belong to the same major topic. The results reinforce the superior performance of the LAM_LEX approach in comparison with other models.

It is worth remarking that in cases where LAM_LEX cluster together sentences labelled with different major topics, some clustering results are actually quite sensible. Table 1 illustrates it with a representative case. These sentences were extracted from a cluster about *farmers* in which 9 out of 10 top topic sentences have "*environment, food and rural affairs*" as the gold major topic. The only discording sentence, belonging to *treasury* (major topic) and *infrastructure investment* (minor topic), is however closely related to *farmers* too and it makes sense to put it into the same cluster.

5.1.2 Perspectiven Summarisation

In this section we evaluate the quality of the relation of the arguments with respect to their topics.

In terms of a quantitative evaluation, we are interested in knowing how strongly the perspectives are related to their topic: it might be the case that the separate CV coherence for the topic and viewpoints is high, but there is no actual relation between them, which would be an undesirable behaviour. To determine whether this is happening or not in the studied models, for each perspective we compute a mixed topic-perspective CV value, by extracting the top 5 perspective words, concatenating them with the top 5 words of the corresponding topic and running Palmetto as in the previous section.[10] We then average the computed mixed topic-perspective CV values by $T \times A$. Following this methodology, a high average CV value means that the perspective words are likely to occur when discussing about that particular topic, and therefore a test of whether the model is learning perspectives that have to do with it. Figure 3 compares topic-perspective models evaluated following this methodology, showing that LAM_LEX gives the best overall coherence.

For a better understanding of what perspec-

[9]The CV results were calculated based on the top 10 words from each topic.

[10]Palmetto does not accept more than 10 words.

Sentence (extracted from a longer speech)	Major topic	Minor topic
I would add that HMRC can provide extra flexibility where there are particular impacts on particular farmers or other businesses	Treasury	Infrastructure Investment
I think milk prices will improve, but the banks need to support farmers in the meantime	Environment food and rural affairs	Topical questions

Table 1: Example sentences, belonging to speeches that were assigned in Hansard different major topics labels, were clustered together by LAM (and it is sensible to do so as they are both about "farmers").

Model	#Topics	≥5	≥6	≥7	≥8	≥9	=10
LDA	10	**0.720**	0.600	0.459	0.320	0.160	0.120
	20	**0.810**	**0.700**	0.570	0.430	0.310	0.180
	30	0.779	0.653	0.580	0.479	0.347	**0.233**
	40	0.620	0.550	0.475	0.360	0.290	0.145
	50	0.732	0.612	0.496	0.388	0.304	0.204
	100	0.654	0.486	0.374	0.306	0.216	0.139
CPT	10	0.580	0.440	0.320	0.260	0.220	0.140
	20	0.530	0.470	0.410	0.340	0.250	0.160
	30	0.473	0.394	0.313	0.273	0.193	0.147
	40	0.420	0.385	0.330	0.250	0.200	0.145
	50	0.464	0.340	0.292	0.220	0.156	0.104
	100	0.435	0.342	0.258	0.190	0.148	0.082
JTV	10	0.620	0.440	0.320	0.199	0.120	0.080
	20	0.634	0.486	0.377	0.303	0.229	0.110
	30	0.753	0.559	0.453	0.319	0.227	0.087
	40	0.705	0.580	0.460	0.370	0.250	0.145
	50	0.628	0.468	0.368	0.290	0.220	0.152
	100	0.636	0.508	0.364	0.274	0.184	0.126
LAM_LEX	10	**0.720**	**0.640**	**0.480**	**0.440**	**0.420**	**0.240**
	20	0.790	0.690	**0.610**	**0.520**	**0.340**	**0.220**
	30	**0.900**	**0.779**	**0.693**	**0.580**	**0.453**	0.213
	40	**0.850**	**0.770**	**0.650**	**0.550**	**0.444**	**0.260**
	50	**0.788**	**0.704**	**0.620**	**0.520**	**0.404**	**0.228**
	100	**0.656**	**0.544**	**0.452**	**0.348**	**0.278**	**0.186**

Table 2: Ratio of topics where x or more than x out of top 10 topic sentences ($\geq x$) belong to the same major topic.

tives LAM_LEX is learning, we extract the top perspective sentences for a given topic based on normalised discriminative score of each sentence[11], similar to what have been done in selecting the top topic sentences. In specific, we first define the discriminative score of word w under topic z and argument a by: $\text{DS}(w,a,z) = \frac{P(w|z,a)}{\max_{z' \neq z, a' \neq a} P(w|z',a')}$. Then the sentence-level discriminative score is calculated based on the aggregated discriminative scores over all the words normalised by the sentence length: $\text{DS}(s,z,a) = \sum_{w \in \boldsymbol{s}} \text{DS}(w,a,z)/\text{Length}(\boldsymbol{s})$. In order to have better correspondence between topics and their respective arguments, we perform two-stage selection: first ranking sentences based on topic-level discriminative scores DS(s,z), and then further ranking sentences based on topic-argument-level discriminative scores $\text{DS}(s,z,a)$.

We can use these extracted top representative sentences together with the gold major topics from HCD to measure if perspectives are connected to their topic. We define the label-based accuracy (LA) as follows: let p_i be the gold major topics associated to the top 10 perspective sentences of a perspective i and let t be the gold major topics corresponding to the top 10 topic sentences; LA(t,p_i) = $\frac{|t \cap p_i|}{|t \cup p_i|}$ measures how many gold major topic labels are shared between topic and perspective sentences. LA also penalises the major topics that are not in common. Table 3 shows for different number of topics the averaged LA measure across all perspectives for three models. It can be observed that LAM_LEX obtains the best performance, followed by CPT.

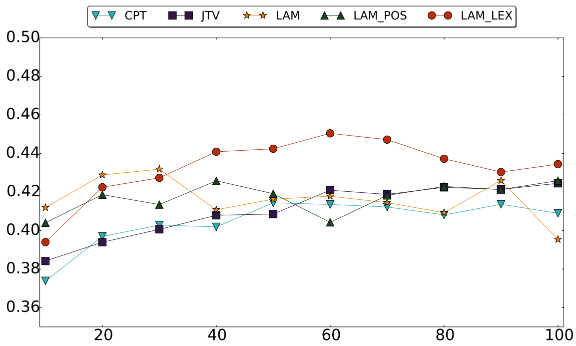

Figure 3: Average mixed topic-perspective CV coherence, across different number of topics.

Topics	CPT	JTV	LAM_LEX
10	0.254	0.308	**0.427**
20	0.369	0.366	**0.517**
30	0.401	0.389	**0.573**
40	0.426	0.394	**0.604**
50	0.431	0.408	**0.564**

Table 3: Averaged LA measure across all topic-perspectives for different models.

To compare the quality of perspectives inferred by LAM_LEX and CPT (over 30 topics) we also conducted human evaluation. To do this, topics and perspectives were represented as bag-of-words. Each perspective was also represented with its three most representative sentences. The outputs from the two models was first merged

[11] We can also rank sentences for an argument a under a topic z based on the generative probability of sentences. But this consistently produce worse results.

and shuffled. Two external annotators were then asked to answer ('yes' or 'no') for each topic if they could differentiate two perspectives. Cohen's Kappa coefficient (Cohen, 1968) for inter-annotator agreement was 0.421. Table 4 shows the results of the evaluation and it is clear that LAM_LEX outperforms CPT.

Annotator	LAM_LEX	CPT
1	**0.63**	0.10
2	**0.67**	0.34
1&2	**0.53**	0.10

Table 4: Accuracy on detecting perspectives according to the human outputs. In 1&2 a 'yes' answer is only valid if marked by both annotators.

Table 5 shows the three most representative perspective sentences for some of the extracted topics by LAM_LEX and CPT, to illustrate how LAM_LEX obtains more coherent sentences.[12] The example involving the first topic shows a case where LAM_LEX learned non-contrastive perspectives: both deal with *Palestina*, but focusing in different aspects (*illegal settlements* vs. *Israel actions*). In contrast, CPT mixed perspectives about Israel/Palestina and other viewpoints about *GCSE* and *mortgages*. In the second topic, LAM_LEX ranks at the top sentences relating to *Sinn Fein & Northern Ireland*, that show two different stances (*positive* vs *negative)* meanwhile in CPT it is not possible to infer any clear perspective despite sentences contain semantically related terms.

Table 6 shows cases where LAM_LEX obtained a less-coherent output according to the annotators. The first topic deals with Shaker Aamer and the legality of its imprisonment in Guantanamo. Perspective 2 reflects this issue, but Perspective 1 includes other types of crimes. The second example discusses issues relating to transports. While Perspective 1 is all about the negotiation with the train company, First Great Western, on its franchise extension proposal, Perspective 2 contains sentences relating to a number of different issues under transports. To alleviate this problem, we hypothesise that additional levels of information (in addition to the topic and perspective levels), such as a Bill or a speaker, might be needed to better distinguish different topics and perspectives that share a significant proportion of vocabulary.

[12]The examples were identified as two perspectives by at least one annotator. Its selection was made based on an existence of a similar topic both on LAM_LEX and CPT outputs.

5.1.3 Discussion

LAM_LEX gave a glimpse of the perspectives that occupy a topic. However, in many cases those differ from the initial expectation given the priors used in our model. Despite of the use of the subjectivity lexicon to initialise the Dirichlet prior β^a for the topic-argument-word distribution, after a few iterations the initial distribution changes radically and turns instead into contrastive and non-contrastive perspectives, with the latter group being the most common one. We think this is due to factors that involve: (1) lack of contrastive speeches about very specific topics; and (2) jargon from the House of Commons that makes the task more challenging as stances are showed in subtle and polite way. This is also in line with what has been previously observed in (Mohammad et al., 2016b) that a person may express the same stance towards a target by using negative or positive language. This shows that LAM_LEX can infer perspectives from raw data, but we have little control on guiding the model on what perspectives to extract.

6 Conclusion and Future Work

We have presented LAM, a model able to provide a glimpse of what is going on in political debates, without relying on any labelled data and assuming the perspectives of a topic to be latent. It is implemented through a hierarchical Bayesian model considering that words can be separated as topic, argument or background words and follow different generative routes. Experiments show that our model obtains more coherent topics than related approaches and also extracts more interpretable perspectives. The code is made available at `https://github.com/aghie/lam`.

Although LAM can extract perspectives under a certain topic, there is little control in what kind of information to extract (e.g. we might want only contrastive or non-contrastive arguments). In future work, we plan to improve the model through complex priors or semantic similarity strategies. Also, adding a 'Bill' level could be beneficial as speeches about the same Bill should share the same high-level topic. But we need labels indicating to which Bill the text belongs to. Including a 'speaker' level to know which parliamentarians discuss which topics is another interesting path to follow.

	LAM_LEX	CPT
Topic 1	israel, iran, syria, settlement, relocation, counter-terrorism gaza, tpims, airline, metropolitan	israel, iran, middle, settlement, palestinian, israeli, gaza, negotiation, village, hamas
Perspective 1	a) It is contrary to international law in that sense, and any nation has obligations when dealing with occupied territories and their occupants. b) Again, I reiterate the difference between the two issues: one concerns the illegal settlements, and the other is a planning matter that we have raised concerns about. c) That is a slightly separate debate or concern if I can put it that way to the illegal settlements that have been put forward, but nevertheless we are concerned and are having a dialogue with Israel about that.	a) Does he agree that unless that happens it is difficult to envisage a unified and prosperous Palestinian state existing alongside Israel? b) Will the Minister discuss that issue with the Israeli Government, urge them to reconsider the upcoming evictions and demolitions due for next month, and instead consider villages co-existing side by side in the spirit of peace? c) That is caused partly by the security situation in Sinai and the Egyptian response to that, and partly by the situation between Israel and the Palestinians in Gaza.
Perspective 2	a) More to the point, the continual encroachment by the Israeli Government makes it impossible for East Jerusalem to become the capital of a Palestinian state. b) We know that 163 Palestinian children are being held in Israeli military detention, and that many are being held inside Israel in direct violation of the fourth Geneva c) We want to see the establishment of a sovereign and independent Palestinian state, living in peace and security alongside Israel.	a) I share the hon. Ladys desire that every school should offer three separate sciences at GCSE; that is very important. b) Everybody here will know, however, that a 1,000 monthly payment sustains a mortgage of 200,000. c) As I clarified, that is a different matter to the debate about the occupied Palestinian territories, but nevertheless we want a robust planning process that adequately.
Topic 2	stormont, sinn, fein, setback, scene, flag, belfast, backwards, surprise, feeling	northern, ireland, stormont, sinn, fein, fairly, poverty, corporation, molyneaux, monday
Perspective 1	a) Would that it was as simple as getting behind the democratic authority in Libyait is not clear that there is a democratic authority behind which we can get. b) It is very important for the Stormont House agreement to be implemented fully and fairly, including all the sections on welfare and budgets. c) The Stormont House agreement was a big step forward, and it is vital for all parties to work to ensure that it is implemented fully and fairly.	a) Following our two major reform programmes, spend has fallen to 1.7 billion in 2013-14 and is expected to fall to about 1.5 billion once the reforms have fully worked through the system. b) Universal credit is a major reform that will transform the welfare state in Britain for the better. c) We have put in place a five-year reform programme that will bring our courts into the 21st century.
Perspective 2	a) There is a clear disparity in political party funding in Northern Ireland, yet Sinn Fein Members continue to draw hundreds of thousands of pounds in allowances from this House, despite not taking their seats. b) In light of the reneging of Sinn Fein on the introduction of welfare reform, what implications does the Minister see in the devolution of corporation tax in Northern Ireland? c) There is no doubt that the announcement by Sinn Fein on Monday was a significant setback for the Stormont House agreement, but it is inevitable that there will be bumps in the road with agreements of this nature.	a) Will the National Crime Agency specifically target the organised criminal gangs that are engaging in subterfuge and in the organised criminal activity of fuel laundering along the border areas of Northern Ireland? b) This will ensure that the people of Northern Ireland are afforded the same protections from serious and organised crime as those in the rest of the United Kingdom. c) The Treasury has had meetings with the European Commission to discuss the reinstatement of the aggregate credit levy scheme for Northern Ireland, which could serve as a further tool of investment in infrastructure.

Table 5: Output sample for representative perspective sentences in non-contrastive and contrastive topics.

	LAM_LEX
Topic 1	aamer, shaker, bay, guantanamo, america, obama, american, timetable, embassy, harlington
Perspective 1	a) NSPCC research has shown that six in 10 teenagers have been asked for sexual images or videos online. b) Does my right hon. Friend agree that the report released last week that suggested that the punishments for online and offline crime should be equalised demonstrates that education is needed to show that the two sentences should be equal? c) I can confirm that the Government have announced that we are entering into a negotiation on a contract for difference for the Swansea bay lagoon to decide whether the project is affordable and represents value for money.
Perspective 2	a) This has been a helpful and constructive debate, and I join others in congratulating the hon. Member for Hayes and Harlington (John McDonnell) on securing it through the Backbench Business Committee. b) I thank the Backbench Business Committee for allocating time for this critical debate at an important time in the campaign to secure the release of Shaker Aamer. c) He has been one of the leading parliamentary campaigners for Mr Aamers release, and I acknowledge the presence of the hon. Member for Battersea (Jane Ellison) , who is the constituency MP for Mr Aamer and his familyindeed, this debate provides an important opportunity to follow up a Backbench Business Committee debate on the same subject that she initiated in April 2013.
Topic 2	passenger, franchise, fare, coast, connectivity, journey, gloucester, user, anglia, stagecoach
Perspective 1	a) Will my hon. Friend confirm when she expects the Departments negotiations with First Great Western on its franchise extension proposals, which include the improvements at Gloucester, to be completed? b) The hon. Gentleman will be pleased to learn that we expect to conclude negotiations with First Great Western and to finalise the second directly awarded franchise contract during this month, and expect the provision of services to start in September. c) My plans for the regeneration of the city of Gloucester include a new car park and entrance to Gloucester station, but they depend on a land sale agreement between the Ministry of Justice and the city council and the lands onward leasing to First Great Western.
Perspective 2	a) I do not want any young people to feel frightened of attending school or of their journey to and from school, and, sadly, that applies particularly to members of the Jewish community at present. b) Why, instead of real localism, have this Government presided over a failed record, with bus fares up 25% and 2,000 routes cut, and a broken bus market, which lets users down, but which Labour will fix in government? c) Last week we introduced the new invitation to tender for the Northern Rail and TransPennine Express services, and transferred East Coast back to the private sector.

Table 6: Output sample for non-representative perspective sentences in the LAM_LEX model.

Acknowledgments

We thank Charles Marshall for crawling the HCD data. DV was funded by MECD (FPU 13/01180), MINECO (FFI2014-51978-C2-2-R) and Inditex-UDC grants for research stays. YH is partly funded by the Natural Science Foundation of China (61528302).

References

Rawia Awadallah, Maya Ramanath, and Gerhard Weikum. 2012. Opinions network for politically controversial topics. In *Proceedings of the first edition workshop on Politics, elections and data*, pages 15–22. ACM.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Filip Boltuzic and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58. Citeseer.

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL): Short Papers-Volume 2*, pages 208–212.

George Casella and Edward I George. 1992. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Sanmay Das and Allen Lavoie. 2014. Automated inference of point of view from user interactions in collective intelligence venues. In *ICML*, pages 82–90.

Yi Fang, Luo Si, Naveen Somasundaram, and Zhengtao Yu. 2012. Mining contrastive opinions on political texts using cross-perspective topic model. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 63–72. ACM.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 375–384. ACM.

Chenghua Lin, Yulan He, Richard Everson, and Stefan Ruger. 2012. Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data engineering*, 24(6):1134–1145.

Marco Lippi and Paolo Torroni. 2016a. Argument mining from speech: Detecting claims in political debates. In *Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*.

Marco Lippi and Paolo Torroni. 2016b. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):10.

Thomas P Minka. 2003. A comparison of numerical optimizers for logistic regression. *Unpublished draft*.

Saif M Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, volume 16.

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2016b. Stance and sentiment in tweets. *arXiv preprint arXiv:1605.01655*.

Huy Nguyen and Diane J Litman. 2015. Extracting argument and domain words for identifying argument components in texts. In *ArgMining@ HLT-NAACL*, pages 22–28.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in mst-style discourse parsing for argumentation mining. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 938–948.

Minghui Qiu, Yanchuan Sim, Noah A Smith, and Jing Jiang. 2015. Modeling user arguments, interactions, and attributes for stance prediction in online debate forums. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 855–863.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408. ACM.

Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument extraction from news. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 56–66.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56.

Amine Trabelsi and Osmar R Zaıane. 2014. Finding arguing expressions of divergent viewpoints in online debates. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)@ EACL*, pages 35–43.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

Guido Zarrella and Amy Marsh. 2016. Mitre at semeval-2016 task 6: Transfer learning for stance detection. *arXiv preprint arXiv:1606.03784*.