A Joint Topic Viewpoint Model for Contention Analysis

Amine Trabelsi and Osmar R. Zaïane

Department of Computing Science, University of Alberta, Edmonton, Canada
{atrabels,zaiane}@ualberta.ca

Abstract. This work proposes an unsupervised Joint Topic Viewpoint model (JTV) with the objective to further improve the quality of opinion mining in contentious text. The conceived JTV is designed to learn the hidden features of arguing expressions. The learning task is geared towards the automatic detection and clustering of these expressions according to the latent topics they confer and the embedded viewpoints they voice. Experiments are conducted on three types of contentious documents: polls, online debates and editorials. Qualitative and quantitative evaluations of the models output confirm the ability of JTV in handling different types of contentious issues. Moreover, analysis of the preliminary experimental results shows the ability of the proposed model to automatically and accurately detect recurrent patterns of arguing expressions.

Keywords: Contention Analysis, Topic Models, Opinion Mining, Unsupervised Clustering.

1 Introduction

Our work fits into the lines of research that addresses the problem of enhancing the quality of opinion extraction from unstructured text found in social media platforms. Netizens use these novel media platforms to discuss and express their opinion over major socio-political events. These events, often, are the object of heated debates over a controversial or contentious issues. A contentious issue is a subject that is likely to stimulate divergent viewpoints within people (e.g., Healthcare Reform, Same-Sex Marriage, Israel/Palestine conflict). In most cases opinion itself is not enough; arguments are needed when people differ on a specific issue. Multiple documents such as surveys' reports, debate sites' posts and editorials may contain multiple contrastive viewpoints regarding a particular issue of contention. Table 1 presents an example of short-text documents expressing divergent opinions where each is exclusively supporting or opposing a healthcare legislation[1]. Opinion in contentious issues is often expressed implicitly, not necessarily through the usage of usual negative or positive opinion words, like "bad" or "great". In addition, the propositional content of the utterances may remain ambiguous in certain circumstances. This makes its extraction a challenging task. Opinion is usually conveyed through the arguing expression justifying the endorsement of a particular point of view. It is advised by the stated words or phrases as they appear in the context. For example, the arguing expression "many people do not have healthcare", in Table 1,

[1] Extracted from a Gallup Inc. survey http://www.gallup.com/poll/126521/favor-oppose-obama-healthcare-plan.aspx

E. Métais, M. Roche, and M. Teisseire (Eds.): NLDB 2014, LNCS 8455, pp. 114–125, 2014.

Table 1. Excerpts of support and opposition opinion to a healthcare bill in the USA

Support Viewpoint	*Oppose Viewpoint*
Many people do not have health care	The government should not be involved
Provide health care for 30 million people	It will produce too much debt
The government should help old people	The bill would not help the people

implicitly explains that the reform is intended to fix the problem of uninsured people, and thus, the opinion is probably on the supporting side. On the other hand, the arguing expression “it will produce too much debt” denotes the negative consequence that may result from passing the bill, making it on the opposing side.

Instead of going through the detailed contents of all documents provided by social media platforms, an automatic concise summary would be appealing for a number of users. For example, it may constitute a rich source of information for policy makers to monitor public opinion and feedback. For journalists, a substantial amount of work can be saved by having automatic access to drafting elements about controversial issues.

The rest of this paper is organized as follows. Section 2 states the problem. Section 3 explains the key issues in the context of recent related work. Section 4 provides the technical details of our model, the Joint Topic Viewpoint model (JTV). Section 5 describes the clustering task that might be used to obtain a feasible solution. Section 6 provides a description of the experimental set up on three different types of contentious text. Section 7 assesses the adequacy and compares the performance of our solution with another model in the literature. Section 8 concludes the paper.

2 Problem Statement

This paper introduces a method of mining important arguing expressions in different types of contentious text (surveys’ reports, debate forums’ posts and editorials). Table 2 presents an example of a human-made summary of arguing expressions [7], obtained from verbatim responses of a survey on the Obama healthcare. Given a corpus of documents, our ultimate goal is to generate similar summaries. However, this paper only concentrates on the subtask of mining the content by first identifying recurrent words and phrases expressing “arguments” and then clustering them according to their topics and viewpoints. Table 2’s examples serve to define key concepts and help formulate the problem. Here, the contentious issue spawning the contrastive viewpoints is the Obama healthcare. The documents are people’s verbatim responses to the question “Why do you favor or oppose a healthcare legislation similar to President Obama’s ?”.

We define a ***contention question*** as a question that can generate expressions of two or more divergent viewpoints as a response.

While the previous question explicitly asks for the reasons (“why”), we relax this constraint and consider also usual opinion questions like “Do you favor or oppose Obamacare ?”, or ”What do you think about Obamacare”.

A ***contentious document*** is a document that contains expressions of one or more divergent viewpoints in response to the contention question.

Table 2. Human-made summary of arguing expressions supporting and opposing Obamacare

Support Viewpoint	***Oppose Viewpoint***
People need health insurance/many uninsured	Will raise cost of insurance/ less affordable
System is broken/needs to be fixed	Does not address real problems
Costs are out of control/help control costs	Need more information on how it works
Moral responsibility to provide/Fair	Against big government involvement (general)
Would make healthcare more affordable	Government should not be involved in healthcare
Don't trust insurance companies	Cost the government too much

Table 2 is split into two parts according to the viewpoint: supporting or opposing the healthcare bill. Each row contains one or more phrases, each expressing a reason (or an explanation), e.g., "System is broken" and "needs to be fixed". Though lexically different, these phrases share a common hidden theme (or topic), e.g., healthcare system, and implicitly convey the same hidden viewpoint's semantics, e.g., support the healthcare bill. Thus, we define an ***arguing expression*** as the set of reasons (snippets: words or phrases) sharing a common topic and justifying the same viewpoint regarding a contentious issue.

We assume that a ***viewpoint*** (e.g., a column of Table 2) in a contentious document is a stance, in response to a contention question, which is implicitly expressed by a set of arguing expressions (e.g., rows of a column in Table 2).

Thus, the arguing expressions voicing the same viewpoint differ in their topics, but agree in the stance. For example, arguing expressions represented by "system is broken" and "costs are out of control" discuss different topics, i.e., healthcare system and insurance's cost, but both support the healthcare bill. On the other hand, arguing expressions of divergent viewpoints may have similar topic or may not. For instance, "government should help elderly" and "government should not be involved" share the same topic "government's role" while conveying opposed viewpoints.

Our research problem and objectives in terms of the newly introduced concepts are stated as follows. Given a corpus of unlabeled contentious documents $\{doc_1, doc_2, .., doc_D\}$, where each document doc_d expresses one or more viewpoints $\boldsymbol{v}^d$ from a set of L possible viewpoints $\{v_1, v_2, .., v_L\}$, and each viewpoint v_l can be conveyed using one or more arguing expressions $\boldsymbol{\phi}_l$ from a set of possible arguing expressions discussing K different topics $\{\phi_{1l}, \phi_{2l}, .., \phi_{Kl}\}$, the objective is to perform the following two tasks:

1. automatically extracting coherent words and phrases describing any distinct arguing expression ϕ_{kl};
2. grouping extracted distinct arguing expressions ϕ_{kl} for different topics, $k = 1..K$, into their corresponding viewpoint v_l.

This paper concentrates on the first task while discussing key elements to realize the second. For the first task, there is a need to account for arguing expressions related to the same topic and viewpoint but having different lexical features, e.g., "provide health care for 30 million people" and "many people do not have healthcare". For this purpose we propose a Joint Topic Viewpoint Model (JTV) to represent the mutual dependence between topics and viewpoints.

3 Related Work

Classifying Stances: An early body of work addresses the challenge of classifying viewpoints in contentious or ideological discourses using supervised techniques [8,10]. Although the models give good performance, they remain data-dependent and costly to label, making the unsupervised approach more appropriate for the existing huge quantity of online data. A similar trend of studies scrutinizes the discourse aspect of a document in order to identify opposed stances [12,16]. However, these methods utilize polarity lexicon to detect opinionated text and do not look for arguing expression, which is shown to be useful in recognizing opposed stances [14]. Somasundaran and Wiebe [14] classify ideological stances in online debates using generated arguing clues from the Multi Perspective Question Answering (MPQA) opinion corpus[2]. Our problem is not to classify documents, but to recognize recurrent pattern of arguing phrases instead of arguing clues. Moreover, our approach is independent of any annotated corpora.

Topic Modeling in Reviews Data: Another emerging body of work applies probabilistic topic models on reviews data to extract appraisal aspects and the corresponding specific sentiment lexicon. These kinds of models are usually referred to as joint sentiment/aspect topic models [6,17,18]. Lin and He [9] propose the Joint Sentiment Topic Model (JST) to model the dependency between sentiment and topics. They make the assumption that topics discussed on a review are conditioned on sentiment polarity. Reversely, our JTV model assumes that a viewpoint endorsement (e.g., oppose reform) is conditioned on the discussed topic (e.g., government's role). Moreover, JTV's application is different from that of JST. Most of the joint aspect sentiment topic models are either semi-supervised or weakly supervised using sentiment polarity words (Paradigm lists) to boost their efficiency. In our case, viewpoints are often expressed implicitly and finding specific arguing lexicon for different stances is a challenging task in itself. Indeed, our model is enclosed in another body of work based on a Topic Model framework to mine divergent viewpoints.

Topic Modeling in Contentious Text: A recent study by Mukherjee and Liu [11] examines mining contention from discussion forums data where the interaction between different authors is pivotal. It attempts to jointly discover contention/agreement indicators (CA-Expressions) and topics using three different Joint Topic Expressions Models (JTE). The JTEs' output is used to discover points (topics) of contention. The model supposes that people express agreement or disagreement through CA-expressions. However, this is not often the case when people express their viewpoint via other channels than discussion forums like debate sites or editorials. Moreover, agreement or disagreement may also be conveyed implicitly through arguing expressions rejecting or supporting another opinion. JTEs do not model viewpoints and use the supervised Maximum Entropy model to detect CA-expressions.

Recently, Gottipati et al. [3] propose a topic model to infer human interpretable text in the domain of issues using Debatepedia[3] as a corpus of evidence. Debatepedia is an online authored encyclopedia to summarize and organize the main arguments of two possible positions. The model takes advantage of the hierarchical structure of arguments

[2] `http://mpqa.cs.pitt.edu/`

[3] `http://dbp.idebate.org`

in Debatepedia. Our work aims to model unstructured online data, with unrestricted number of positions, in order to, ultimately, help extract a relevant contention summary.

The closest work to ours is the one presented by Paul et al. [13]. It introduces the problem of contrastive summarization which is very similar to our stated problem in Section 2. They propose the Topic Aspect Model (TAM) and use the output distributions to compute similarities' scores for sentences. Scored sentences are used in a modified Random Walk algorithm to generate the summary. The assumption of TAM is that any word in the document can exclusively belong to a topic (e.g., government), a viewpoint (e.g., good), both (e.g., involvement) or neither (e.g., think). However, according to TAM's generative model, an author would choose his viewpoint and the topic to talk about independently. Our JTV encodes the dependency between topics and viewpoints.

4 Joint Topic Viewpoint Model

Latent Dirichlet Allocation (LDA) [2] is one of the most popular topic models used to mine large text data sets. It models a document as a mixture of topics where each topic is a distribution over words. However, it fails to model more complex structures of texts like contention where viewpoints are hidden. We augment LDA to model a contentious document as a pair of dependent mixtures: a mixture of arguing topics and a mixture of viewpoints for each topic. The assumption is that a document discusses the topics in proportions, (e.g., 80% government's role, 20% insurance's cost). Moreover, as explained in Section 2, each one of these topics can be shared by divergent arguing expressions conveying different viewpoints. We suppose that for each discussed topic in the document, the viewpoints are expressed in proportions. For instance, 70% of the document's text discussing the government's role expresses an opposing viewpoint to the reform while 30% of it conveys a supporting viewpoint. Thus, each term in a document is assigned a pair topic-viewpoint label (e.g., "government's role-oppose reform"). A term is a word or a phrase i.e., n-grams (n>1). For each topic-viewpoint pair, the model generates a topic-viewpoint probability distribution over terms. This topic-viewpoint distribution would correspond to what we define as an arguing expression in Section 2, i.e., a set of terms sharing a common topic and justifying the same viewpoint regarding a contentious issue.

Formally, assume that a corpus contains D documents $d_{1..D}$, where each document is a term's vector $\boldsymbol{w}_d$ of size N_d; each term w_{dn} in a document belongs to the corpus vocabulary of distinct terms of size V. Let K be the total number of topics and L be the total number of viewpoints. Let θ_d denote the probabilities (proportions) of K topics under a document d; ψ_{dk} be the probability distributions (proportions) of L viewpoints for a topic k in the document d (the number of viewpoints L is the same for all topics); and ϕ_{kl} be the multinomial probability distribution over terms associated with a topic k and a viewpoint l. The generative process (see. the JTV graphical model in Fig. 1) is:

- for each topic k and viewpoint l, draw a multinomial distribution over the vocabulary V: $\phi_{kl} \sim Dir(\beta)$;
- for each document d,
  - draw a topic mixture $\theta_d \sim Dir(\alpha)$
  - for each topic k, draw a viewpoint mixture $\psi_{dk} \sim Dir(\gamma)$

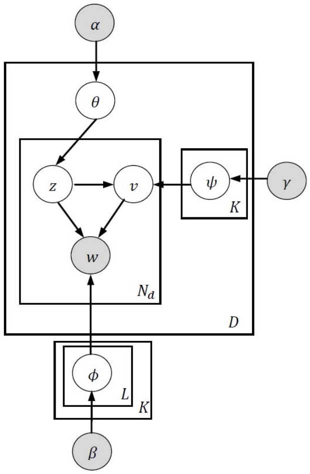


Fig. 1. The JTV's graphical model (plate notation)

- for each term w_{dn}
  - $*$ sample a topic assignment $z_{dn} \sim Mult(\theta_d)$
  - $*$ sample a viewpoint assignment $v_{dn} \sim Mult(\psi_{dz_{dn}})$
  - $*$ sample a term $w_{dn} \sim Mult(\phi_{z_{dn}v_{dn}})$

We use fixed symmetric Dirichlet's parameters γ, β and α. They can be interpreted as the prior counts of: terms assigned to viewpoint l and topic k in a document; a particular term w assigned to topic k and viewpoint l within the corpus; terms assigned to a topic k in a document, respectively. In order to learn the hidden JTV's parameters ϕ_{kl}, ψ_{dk} and θ_d, we draw on approximate inference as exact inference is intractable [2]. We use the collapsed Gibbs Sampling [4], a Markov Chain Monte Carlo algorithm. The collapsed Gibbs sampler integrate out all parameters ϕ, ψ and θ in the joint distribution of the model and converge to a stationary posterior distribution over viewpoints' assignments $\boldsymbol{v}$ and all topics' assignments $\boldsymbol{z}$ in the corpus. It iterates on each current observed token w_i and samples each corresponding v_i and z_i given all the previous sampled assignments in the model $\boldsymbol{v}_{\neg i}$, $\boldsymbol{z}_{\neg i}$ and observed $\boldsymbol{w}_{\neg i}$, where $\boldsymbol{v} = \{v_i, \boldsymbol{v}_{\neg i}\}$, $\boldsymbol{z} = \{z_i, \boldsymbol{z}_{\neg i}\}$, and $\boldsymbol{w} = \{w_i, \boldsymbol{w}_{\neg i}\}$. The derived sampling equation is:

$$p(z_i = k, v_i = l | \boldsymbol{z}_{\neg i}, \boldsymbol{v}_{\neg i}, w_i = t, \boldsymbol{w}_{\neg i}) \propto \frac{n^{(t)}_{kl,\neg i} + \beta}{\sum_{t=1}^{V} n^{(t)}_{kl,\neg i} + V\beta} . \frac{n^{(l)}_{dk,\neg i} + \gamma}{\sum_{l=1}^{L} n^{(l)}_{dk,\neg i} + L\gamma} . n^{(k)}_{d,\neg i} + \alpha \quad (1)$$

where $n^{(t)}_{kl,\neg i}$ is the number of times term t was assigned to topic k and the viewpoint l in the corpus; $n^{(l)}_{dk,\neg i}$ is the number of times viewpoint l of topic k was observed in document d; and $n^{(k)}_{d,\neg i}$ is the number of times topic k was observed in document d.

All these counts are computed excluding the current token i, which is indicated by the symbol $\neg i$. After the convergence of the Gibbs algorithm, the parameters ϕ, ψ and θ are estimated using the last obtained sample. The probability that a term t belongs to a viewpoint l of topic k is approximated by:

$$\phi_{klt} = \frac{n_{kl}^{(t)} + \beta}{\sum\limits_{t=1}^{V} n_{kl}^{(t)} + V\beta}. \tag{2}$$

The probability of a viewpoint l of a topic k under document d is estimated by:

$$\psi_{dkl} = \frac{n_{dk}^{(l)} + \gamma}{\sum\limits_{l=1}^{L} n_{dk}^{(l)} + L\gamma}. \tag{3}$$

The probability of a topic k under document d is estimated by:

$$\theta_{dk} = \frac{n_{d}^{(k)} + \alpha}{\sum\limits_{k=1}^{K} n_{d}^{(k)} + K\alpha}. \tag{4}$$

5 Clustering Arguing Expressions

Although we are not tackling the task of clustering arguing expressions according to their viewpoints in this paper (Task 2 in Section 2), we explain how the structure of JTV lays the ground for performing it. We mentioned in the previous Section that an inferred topic-viewpoint distribution ϕ_{kl} can be assimilated to an arguing expression. For convenience, we will use “arguing expression” and “topic-viewpoint” interchangeably to refer to the topic-viewpoint distribution. Indeed, two topic-viewpoint ϕ_{kl} and $\phi_{k'l}$, having different topics k and k', do not necessarily express the same viewpoint, despite the fact that they both have the same index l. The reason stems from the nested structure of the model, where the generation of the viewpoint assignments for a particular topic k is completely independent from that of topic k'. In other words, the model does not trace and match the viewpoint labeling along different topics. Nevertheless, the JTV can still help overcome this problem. According to the JTV's structure, a topic-viewpoint ϕ_{kl}, is more similar in distribution to a divergent topic-viewpoint $\phi_{kl'}$, related to the same topic k, than to any other topic-viewpoint $\phi_{k'*}$, corresponding to a different topic k'. Therefore, we can formulate the problem of clustering arguments as a constrained clustering problem [1]. The goal is to group the similar topics-viewpoints ϕ_{kl}s into L clusters (number of viewpoints), given the constraint that the ϕ_{kl}s of the same topic k should not belong to the same cluster.

6 Experimental Set Up

In order to evaluate the performances of the JTV model, we utilize three types of multiple contrastive viewpoint text data: (1) short-text data where people express their viewpoint briefly with few words like survey's verbatim response or social media posts; (2)

Table 3. Statistics on the three used data sets

	GM		IP		OC	
Viewpoint	hurt	no	pal	is	for	ag
#doc	149	301	149	149	434	508
total #toks	47915		209481		14594	
avg. #toks per doc	106.47		702.95		15.94	

mid-range text where people develop their opinion further using few sentences, usually showcasing illustrative examples justifying their stances; (3) long text data, mainly editorials where opinion is expressed in structured and verbose manner.

Throughout the evaluation procedure, analysis is performed on three different types of data sets, corresponding to three different contention issues. Table 3 describes the used data sets. **ObamaCare (OC)**[4] consists of short verbatim responses concerning the "Obamacare" bill. The survey was conducted by Gallup®from March 4-7, 2010. People were asked why they would oppose or support a bill similar to Obamacare. Table 2 is a human-made summary of this corpus. **Gay Marriage (GM)**[5] contains posts in "createdebate.com" responding to the contention question "How can gay marriage hurt anyone?". Users indicate the stance of their posts (i.e., "hurts everyone? (does hurt)" or "doesn't hurt"). **Israel-Palestine (IP)**[6] data set is extracted from BitterLemons web site. It contains articles of two permanent editors, a Palestinian and an Israeli, about the same issue. Articles are published weekly from 2001 to 2005. They discuss several contention issues, e.g., "the American role in the region" and "the Palestinian election".

Paul et al. [13] stress out the importance of negation features in detecting contrastive viewpoints. Thus, we performed a simple treatment of merging any negation indicators, like "nothing", "no one", "never", etc., found in text with the following occurring word to form a single token. Moreover, we merge the negation "not" with any auxiliary verb (e.g., is, was, could, will) preceding it. Then, we removed the stop-words.

Throughout the experiments below, the JTV's hyperparameters are set to fixed values. The γ is set, according to Steyvers and Griffiths's [15] hyperparameters settings, to $50/L$, where L is the number of viewpoints. β and α are adjusted manually, to give reasonable results, and are both set to 0.01. Along the experiments, we try different number of topics K. The number of viewpoints L is equal to 2. The TAM model [13] (Section 3) is run as a means of comparison during the evaluation (with default parameters).

7 Model Evaluation

7.1 Qualitative Evaluation

We perform a qualitative analysis of JTV using the ObamaCare data set. Tables 4 presents the inferred topic-viewpoints, i.e., arguing expressions. We set a number of

[4] http://www.gallup.com/poll/126521/favor-oppose-obama-healthcare-plan.aspx

[5] http://www.createdebate.com/debate/show/How_can_gay_marriage_hurt_any_one

[6] http://www.bitterlemons.net/

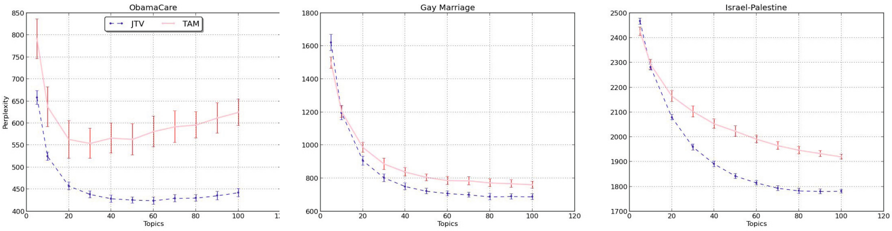
Table 4. JTV's generated topics-viewpoints from Obamacare data set

Topic 1 0.19		*Topic 2 0.20*		*Topic 3 0.20*	
View 1 0.55	*View 2 0.45*	*View 3 0.51*	*View 4 0.49*	*View 5 0.54*	*View 6 0.46*
healthcare	dont_think	people	government	insurance	country
system	work	cant_afford	dont_want	health	economy
uninsured	bill	doctors	involved	companies	medicine
country	abortion	lack	control	years	dollars
world	fair	covered	dont_think	prices	american
change	debt	americans	dont_like	reason	start

Topic 4 0.21		*Topic 5 0.20*	
View 7 0.55	*View 8 0.45*	*View 9 0.47*	*View 10 0.53*
healthcare	healthcare	people	costs
cost	cost	money	medicare
expensive	coverage	pay	increase
afford	dont_know	dont_have	pay
care	public	children	worse
feel	preexisting	poor	problems

topics of $K = 5$ and a number of viewpoints of $L = 2$. Each topic-viewpoint (e.g., Topic 1-View 1) is represented by the set of top terms. The terms are sorted in descending order according to their probabilities. Inferred probabilities over topics, and over viewpoints for each topic, are also reported. We try to qualitatively observe the distinctiveness of each arguing (topic-viewpoint) and assess its coherence in terms of the topic discussed and the viewpoint conveyed and its divergence with the opposing pair-element. In Table 4 most of the topic-viewpoint pairs, corresponding to the same topic, are conveying opposite stances. For instance, taking a closer look at the original data suggests that Topic 1-View 1 (Table 4) criticizes the healthcare system and stresses out the need for a change (e.g., " *We ought to **change** the **system** so everyone can have it (the healthcare insurance)*", "*Because the greatest **country** in the **world** has a dismal **healthcare system***"). This may correspond to the second support arguing expressions in the reference summary of Table 2. On the other side, Topic 1-View 2 may convey the belief that the bill will not work or that it is not fair e.g., "I **don't think** it's **fair**". It also opposes the bill for including the abortion and for the debt that it may induce. Although the debt and the abortion are not related, as topics, they both tend to be adduced by people opposing the bill. Similarly, Topic 2-View 3 may reveal that people can't afford healthcare and they need to be covered (first support arguing in Table 2). However, the opposite side seems to be not enthusiastic about the government's involvement and control (fourth and fifth oppose arguing expressions in Table 2). The same pattern is observed in Topic 3. A matching can also be established with the reference summary.

Detecting different arguing expressions for the same topic proves to be a difficult task when the reasons are lexically very similar. An example is Topic 4 in Table 4 where the shared topic is "healthcare cost". In this case, both arguing expressions are about high costs. The original data contains two rhetoric: one is about current existing costs (supporting side) and the other is about costs induced by the bill (opposing side). However, both Topic 4-View 7 and Topic 4-View 8 seem to convey the supporting viewpoint. The increasing costs yield by the bill may be conveyed in Topic 5-View 10.



(a) OC (b) GM (c) IP

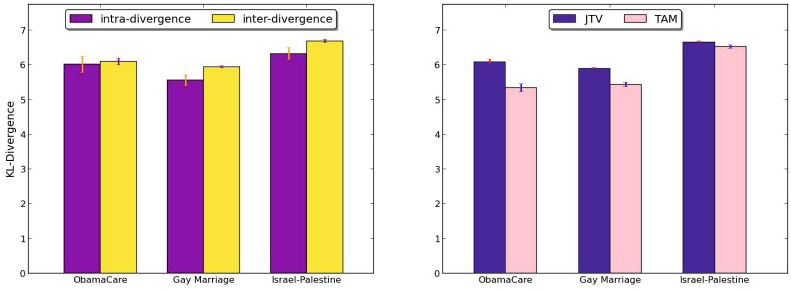
Fig. 2. JVT and TAM's perplexity plots for three different data sets

7.2 Quantitative Evaluation

We assess the ability of the model to fit three data sets and to generate distinct topic-viewpoint by comparing it with TAM which also models the topic-viewpoint dimension.

Held-Out Perplexity. We use the perplexity criterion to measure the ability of the learned topic model to fit a new held-out data. Perplexity assesses the generalization performance and, subsequently, provides a comparing framework of learned topic models. The lower the perplexity, the less "perplexed" is the model by unseen data and the better the generalization. It algebraically corresponds to the inverse geometrical mean of the test corpus' terms likelihoods given the learned model parameters [5]. We compute the perplexity under estimated parameters of JTV and compare it to that of TAM for our three unigrams data sets (Section 6). Figure 2 exhibits, for each corpus, the perplexity plot as function of the number of topics K for JTV and TAM. Note that for each K, we run the model 50 times. The drawn perplexity corresponds to the average perplexity on the 50 runs where each run compute one-fold perplexity from a 10-fold cross-validation. The figures show evidence that the JTV outperforms TAM for all data sets, used in the experimentation.

Kullback-Leibler Divergence. Kullback-Leibler (KL) Divergence is used to measure the degree of separation between two probability distributions. We utilize it for two purposes. The first purpose is to validate the assumption we made in Section 5 which states that, according to JTV's structure, a topic-viewpoint ϕ_{kl} is more similar in distribution to a topic-viewpoint $\phi_{kl'}$, related to the same topic k, than to any other topic-viewpoint $\phi_{k'*}$, corresponding to a different topic k'. Thus, two measures of *intra* and *inter-divergence* are computed. The *intra-divergence* is an average KL-Divergence between all topic-viewpoint distributions that are associated with a same topic. The *inter-divergence* is an average KL-Divergence between all pairs of topic-viewpoint distributions belonging to different topics. Figure 3a displays the histograms of JTV's intra and inter divergence values for the three data sets. These quantities are averages on 20 runs of the model for an input number of topics $K = 5$, which gives the best differences between the two measures. We observe that a higher divergence is recorded between



(a)

(b)

Fig. 3. Histograms of: (a) average topic-viewpoint intra/inter divergences of JTV; (b) average of overall topic-viewpoint divergences of JTV and TAM ($K = 5$)

topic-viewpoints of different topics than between those of a same topic. This is verified for all the data sets considered in our experimentation. The differences between the intra and inter divergences are significant ($p - value < 0.01$) over unpaired t-test (except for Obamacare). The second purpose of using KL-Divergence is to assess the distinctiveness of generated topic-viewpoint by JTV and TAM. This is an indicator of a good aggregation of arguing expressions. We compute an *overall-divergence* quantity, which is an average KL-Divergence between all pairs of topic-viewpoint distributions, for JTV and TAM and compare them. Figure 3b illustrates the results for all datasets. Quantities are averages on 20 runs of the models. Both models are run with a number of topics $K = 5$, which gives the best divergences for TAM. Comparing JTV and TAM, we notice that the overall-divergence of JTV's topic-viewpoint is significantly ($p - value < 0.01$) higher for all data sets. This result reveals a better quality of our JTV extracting process of arguing expressions (the first task stated in Section 2).

8 Conclusion and Future Work

Within the framework of probabilistic graphical models, we presented an approach to mining the important topics and divergent viewpoints in contentious opinionated text. We proposed a Joint Topic Viewpoint model (JTV) for the unsupervised detection and clustering of recurrent arguing expressions. Preliminary results show that our model can provide accommodation for various types of texts (survey reports, debate forums posts and editorials). Moreover, the detection and clustering accuracy has been shown to be enhanced by accounting for mutual dependence of topics and viewpoints. Future work study needs to improve the topicality coherence of extracted arguing phrases. It should also give more insights into their clustering according to their viewpoints, as well as their automatic extractive summary. A human-oriented evaluation of generated arguing expressions and summaries needs to be set up.

References

1. Basu, S., Davidson, I., Wagstaff, K.: Constrained Clustering: Advances in Algorithms, Theory, and Applications, 1st edn. Chapman & Hall/CRC (2008)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
3. Gottipati, S., Qiu, M., Sim, Y., Jiang, J., Smith, N.A.: Learning topics and positions from debatepedia. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (2013)
4. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America 101(1), 5228–5235 (2004)
5. Heinrich, G.: Parameter estimation for text analysis. Tech. rep., Fraunhofer IGD (September 2009)
6. Jo, Y., Oh, A.H.: Aspect and sentiment unification model for online review analysis. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 815–824 (2011)
7. Jones, J.M.: In u.s., 45% favor, 48% oppose obama healthcare plan (March 2010), `http://www.gallup.com/poll/126521/favor-oppose-obama-healthcare-plan.aspx`
8. Kim, S.M., Hovy, E.H.: Crystal: Analyzing predictive opinions on the web. In: Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1056–1064 (2007)
9. Lin, C., He, Y.: Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 375–384 (2009)
10. 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 the Tenth Conference on Computational Natural Language Learning, pp. 109–116 (2006)
11. Mukherjee, A., Liu, B.: Mining contentions from discussions and debates. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 841–849 (2012)
12. Park, S., Lee, K., Song, J.: Contrasting opposing views of news articles on contentious issues. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 340–349 (2011)
13. Paul, M.J., Zhai, C., Girju, R.: Summarizing contrastive viewpoints in opinionated text. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 66–76 (2010)
14. Somasundaran, S., Wiebe, J.: Recognizing stances in ideological on-line debates. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 116–124 (2010)
15. Steyvers, M., Griffiths, T.: Probabilistic topic models. In: Handbook of Latent Semantic Analysis, vol. 427(7), pp. 424–440 (2007)
16. Thomas, M., Pang, B., Lee, L.: Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp. 327–335 (2006)
17. Titov, I., McDonald, R.: Modeling online reviews with multi-grain topic models. In: Proceedings of the 17th International Conference on World Wide Web, pp. 111–120 (2008)
18. Zhao, W.X., Jiang, J., Yan, H., Li, X.: Jointly modeling aspects and opinions with a maxent-lda hybrid. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 56–65 (2010)