

Mining Contentious Documents Using an Unsupervised Topic Model Based Approach

Amine Trabelsi and Osmar R. Zaiane
Department of Computing Science
University of Alberta
Edmonton, Canada
Email: {atrabels,zaiane}@ualberta.ca

Abstract—This work proposes an unsupervised method intended to enhance the quality of opinion mining in contentious text. It presents a Joint Topic Viewpoint (JTV) probabilistic model to analyse the underlying divergent arguing expressions that may be present in a collection of contentious documents. It extends the original Latent Dirichlet Allocation (LDA), which makes it domain and thesaurus-independent, e.g., does not rely on WordNet coverage. The conceived JTV has the potential of automatically carrying the tasks of extracting associated terms denoting an arguing expression, according to the hidden topics it discusses and the embedded viewpoint it voices. Furthermore, JTV’s structure enables the unsupervised grouping of obtained arguing expressions according to their viewpoints, using a constrained clustering approach. Experiments are conducted on three types of contentious documents: polls, online debates and editorials. The qualitative and quantitative analysis of the experimental results show the effectiveness of our model to handle six different contentious issues when compared to a state-of-the-art method. Moreover, the ability to automatically generate distinctive and informative patterns of arguing expressions is demonstrated.

I. INTRODUCTION

Sentiment analysis, also referred to as opinion mining, is an active research area in natural language processing as well as data mining, that aims to extract and examine opinions, attitudes and emotions expressed in text, with respect to some topic in blog posts, comments and reviews. In addition to sentiment expressed towards products, other online text sources such as opinion polls, debate websites and editorials may contain valuable opinion information articulated around contentious topics. In this paper, we address the issue of improving the quality of opinion mining from contentious texts, found in surveys’ responses, debate websites and editorials. Mining and summarizing these resources is crucial, especially when the opinion is related to a subject that stimulates divergent viewpoints within people (e.g., Healthcare Reform, Same-Sex Marriage, Israel/Palestine conflict). We refer to such subjects as issues of contention. A *contentious issue* is “likely to cause disagreement between people” (cf. Oxford Dictionaries¹). Documents such as survey reports, debate site posts and editorials may contain multiple contrastive viewpoints regarding a particular issue of contention. Table I presents an example of short-text documents expressing divergent opinions where each is

TABLE I. EXCERPTS OF SUPPORT AND OPPOSITION OPINION TO A HEALTHCARE BILL IN THE USA.

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

exclusively supporting or opposing a healthcare legislation². Opinion in contentious issues is often expressed implicitly, not necessarily through the usage of usual negative or positive opinion words, like “bad” or “great”. This makes its extraction a challenging task. It is usually conveyed through the arguing expression justifying the endorsement of a particular point of view. The act of arguing is “to give reasons why you think that something is right/wrong, true/not true, etc, especially to persuade people that you are right” (cf. Oxford Dictionaries). For example, the arguing expression “many people do not have healthcare”, in Table I, 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.

The automatic identification and clustering of these kind of arguing expressions, according to the topics they invoke and the viewpoints they convey, is enticing for a variety of application domains. For instance, it can save journalists a substantial amount of work and provide them with drafting elements (viewpoints and associated arguing expressions) about controversial issues. Moreover, a good automatic browsing of divergent arguing expressions in a conflict/issue would help inquisitive people understand the issue itself (e.g., same-sex marriage). Also, it may be used by politicians to monitor the change in argumentation trends, i.e., changes in the main reasons expressed to oppose or support viewpoints. The significant changes may indicate the occurrence of an important event (e.g., a success of a politician’s action or speech). Automatic summarization of arguing expressions may benefit survey companies who usually collect large verbatim reports

¹<http://www.oxfordlearnersdictionaries.com/definition/english/contentious>

²extracted from a Gallup Inc. survey <http://www.gallup.com/poll/126521/favor-oppose-obama-healthcare-plan.aspx>

about people’s opinion regarding an issue of contention. From a text mining perspective, representing a document describing or containing a contention, as a set of arguing expressions from different viewpoints, is useful for information retrieval tasks like query answering or dimensionality reduction. In addition, it would enhance the output quality of the opinion summarization task in general.

The rest of this paper is organized as follows. Section II states the problem. Section III explains the key issues in the context of recent related work. Section IV provides the technical details of the proposed model, the Joint Topic Viewpoint model (JTV). Section V describes the clustering task used to obtain a feasible solution. Section VI provides a description of the experimental set up on three different types of contentious text. Section VII assesses the adequacy and compares the performance of our solution with another model in the literature. Section VIII discusses the future work.

II. PROBLEM STATEMENT

This paper examines the task of mining the underlying topics and the hidden viewpoints of arguing expressions towards the summarization of contentious text. An example of a human-made summary of arguing expressions [1] on, what is commonly known as the Obama healthcare reform, is presented in Table II. The ultimate research’s target is to automatically generate similar snippet-based summaries given a corpus of contentious documents. However, this paper tackles the initial sub-problem of identifying recurrent words and phrases expressing arguing and cluster them according to their topics and viewpoints. This would help solve the general problem. Indeed, the clustered words and phrases can be used as input to query the original documents via information retrieval methods in order to extract relevant fragments or snippets of text related to a particular arguing expression. We use Table II examples to define some key concepts which can help us formulate the general problem. Here, the contentious issue yielding the divergent positions 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?”.

A **contention question** is 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 a contention question.

Table II 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.

a **viewpoint** (e.g., a column of Table II) 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 II).

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 $\{doc_1, doc_2, \dots, doc_D\}$, where each document doc_d expresses one or more viewpoints \vec{v}^d from a set of L possible viewpoints $\{v_1, v_2, \dots, v_L\}$, and each viewpoint v_l can be conveyed using one or more arguing expressions $\vec{\phi}_l$ from a set of possible arguing expressions discussing K different topics $\{\phi_{1l}, \phi_{2l}, \dots, \phi_{Kl}\}$, 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 .

In carrying out the first task, we must meet the main challenge of recognizing arguing expressions having the same topic and viewpoint but which are lexically different, 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 account for the dependence structure of topics and viewpoints. For the second task, the challenge is to deal with the situation where an arguing expression, associated with a specific topic, may share more common words and phrases with a divergent argument, discussing the same topic, than with another argument conveying the same viewpoint but discussing a different topic. Recall, the example “government should help elderly” is lexically more similar to “government should not be involved” than to “many people uninsured”.

III. RELATED WORK

It is important to note that we do not intend to address argumentation analysis. A large body of early work on argumentation was based on learning deterministic logical concepts [2]. Argumentation theory is the study of how conclusions can be reached from some premises through logical reasoning. In argumentation, one critically examines beliefs to discard wrong claims and build knowledge from supported assertions following the Cartesian view of reasoning. In this work, our targeted text is online text in opinion polls, discussion forums, etc. voicing opinions of laypersons. Apart from long editorials, these text sources are typically short in which reasoning is

TABLE II. HUMAN-MADE SUMMARY OF ARGUING EXPRESSIONS SUPPORTING AND OPPOSING OBAMACARE.

<i>Support Viewpoint</i>	<i>Oppose Viewpoint</i>
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

not necessarily laid out but claims and point of views are put forward using arguing expressions. There is little or no rationalization or discursive reasoning in online short surveys or micro blogs. Moreover dealing with these types of opinionated real data, unavoidably requires the means to handle the uncertainty (as opposed to determinism) or the ambiguity that arises from incomplete or hidden information (implicit, unsaid or unexpressed topic or a viewpoint). Our objective is to design a statistical learning model in order to discover related arguing expressions and group them by viewpoint. In this section we present a number of the common themes, issues and important concepts in some related work. Potential links to our approach of mining opinion in text of contention are put forward.

Classifying Stances: An early body of work addresses the challenge of classifying viewpoints in contentious or ideological discourses using supervised techniques [3], [4]. 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 [5], [6]. 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 [7]. Somasundaran and Wiebe [7] classify ideological stances in online debates using generated arguing clues from the Multi Perspective Question Answering (MPQA) opinion corpus³. 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 [8], [9], [10]. Lin and He [11] 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: Lin et al. [12] propose a probabilistic graphical model for ideological discourse. This model takes into account lexical variations between authors having different ideological perspectives. The authors empirically show its effectiveness in fitting ideological texts. However, their model assumes that the perspectives expressed in the documents are observed, while, in our work, the viewpoint labels of the contentious documents are hidden.

A recent studies by Mukherjee and Liu [13], [14] examine mining contention from discussion forums data where the interaction between different authors is pivotal. They attempt 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.

Qiu and Jiang [15] also incorporate the information on users interactions in threaded debate forums within a topic model. The goal is to model both the posts and the users in a thread and cluster them according to their viewpoints. The topic model is based on three major hypothesis: (1) the topics discussed in divergent viewpoints tend to be different; (2) a user is holding the same viewpoint in all his posts in the thread; and (3) users with the same viewpoints have positive interactions while negative interactions are more probable in the opposite case. In our work, we assume that topics are shared between divergent viewpoints. However, the topics' proportions and their related lexicon are different according to the viewpoint. We focus on capturing the lexical variations between divergent viewpoints, instead of the agreement/disagreement between users. While the users interactions can be very useful for posts classification or clustering, our primary goal is different, i.e., it aims at extracting and clustering meaningful arguing expressions towards the summarization of main contention points in an issue. Moreover, our model tends to be generalizable to different types of contentious text (e.g., surveys responses, editorials) which do not necessarily embrace the same structure of threaded debate forums (i.e., do not contain users information and users interaction).

Fang et al. [16] proposed a Cross-Perspective Topic model (CPT) that takes as input separate collections in the political domain, each related to particular viewpoint (perspective). It finds the shared topics between these different collections and the opinion words corresponding to each topic in a collection. However, CPT does not model the viewpoint variable. Thus,

³<http://mpqa.cs.pitt.edu/>

it cannot cluster documents according to their viewpoints. Moreover, the discovered topics are not necessarily of contention. Recently, Gottipati et al. [17] propose a topic model to infer human interpretable text in the domain of issues using Debatepedia⁴ 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 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. [18]. It introduces the problem of contrastive summarization which is very similar to our stated problem in Section II. 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.

IV. JOINT TOPIC VIEWPOINT MODEL

Latent Dirichlet Allocation (LDA) [19] 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 II, 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 II, i.e., a set of terms sharing a common topic and justifying the same viewpoint regarding a contentious issue.

Formally, we assume that a corpus contains D documents $d_{1..D}$, where each document is a term's vector \vec{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)

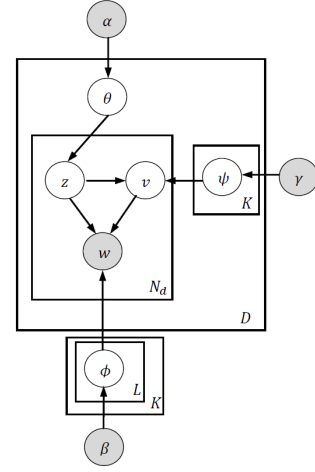


Fig. 1. The JTV's graphical model (plate notation)

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 $\psi_{dk} \sim Dir(\gamma)$
 - 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 [19]. We use the collapsed Gibbs Sampling [20], 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 \vec{v} and all topics' assignments \vec{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 \vec{v}_{-i} , \vec{z}_{-i} and observed \vec{w}_{-i} , where $\vec{v} = \{v_i, \vec{v}_{-i}\}$, $\vec{z} = \{z_i, \vec{z}_{-i}\}$, and $\vec{w} = \{w_i, \vec{w}_{-i}\}$. The derived sampling equation is:

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

⁴<http://dbp.idebate.org>

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

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

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

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

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

$$\theta_{dk} = \frac{n_d^{(k)} + \alpha}{\sum_{k=1}^K n_d^{(k)} + K\alpha}. \quad (4)$$

V. CLUSTERING ARGUING EXPRESSIONS

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 probably 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' (we verify this assumption in Section VII-B2). Therefore, we can formulate the problem of clustering arguing expressions as a constrained clustering problem [21]. The goal is to group the similar topics-viewpoints ϕ_{kls} into L clusters (number of viewpoints), given the constraint that the L ϕ_{kls} of the same topic k should not belong to the same cluster (cannot-link constraints). Thus, each cluster C_i where $i = 1..L$ will contain exactly K topics-viewpoints.

We suggest a slightly modified version of the constrained k-means clustering (COP-KMEANS) [22]. It is presented in Algorithm 1. Unlike COP-KMEANS, we do not consider any must-link constraint but only the above mentioned cannot-link constraints. The centers of clusters are initialized with the topic-viewpoint distributions of the most frequent topic

Algorithm 1 Topic-Viewpoint Clustering

Require: JTV’s output:topic-viewpoint distributions ϕ_{kls} , number of topics K , number of viewpoints L

- 1: Initialize the set C with a set of empty clusters; Choose the topic-viewpoint distributions $\phi_{k^\dagger 1} \dots \phi_{k^\dagger L}$ of the most frequent topic k^\dagger according to JTV as the initial cluster centers.
- 2: **for** each topic k ($k = 1..K$) **do**
- 3: F (clusters to fill) is a copy of set C
- 4: A is a set of L topic-viewpoints ϕ_{kl} to assign (having the same topic k)
- 5: **while** F is not empty **do**
- 6: **for** each ϕ_{kl} in A **do**
- 7: find the closest C_i in F
- 8: add ϕ_{kl} to potential cluster assignment set S_i (corresponding to cluster C_i)
- 9: **end for**
- 10: **for** each cluster C_i **do**
- 11: **if** the corresponding S_i is not empty **then**
- 12: find ϕ_{kl}^* in S_i with the minimum distance from C_i ’s center and assign it to C_i .
- 13: Update C
- 14: empty S_i
- 15: remove ϕ_{kl}^* from A /remove C_i from F
- 16: **end if**
- 17: **end for**
- 18: **end while**
- 19: **end for**
- 20: Update each cluster C_i ’s center by averaging all $\phi^{(i)}$ that have been assigned to it.
- 21: Repeat 2 to 20 until convergence
- 22: **return** set of clusters C

k^\dagger according to the output of JTV. The idea is that it is more probable to find at least one most frequent topic-viewpoint pair for a viewpoint l in the most frequent topic k^\dagger . The cannot-link constraints are implicitly coded in Algorithm 1. Indeed, we constrain the set of L topic-viewpoint ϕ_{kls} of the same topic k (line 2 to 18) to be in a one-to-one matching with the set C of L clusters (lines 5 to 18). Iteratively, the best match, producing a minimal distance between unassigned topic-viewpoints (of the same topic) and the remaining available clusters, is first established (lines 10 to 16). The distance between a topic-viewpoint distribution ϕ_{kl} and another distribution ϕ_* is measured using the symmetric Jensen-Shannon Distance (D_{JS}) [23] which is based on the Kullback-Leibler Divergence (D_{KL}) [24]:

$$D_{JS}(\phi_{kl}||\phi_*) = \frac{1}{2}[D_{KL}(\phi_{kl}||M) + D_{KL}(\phi_*||M)], \quad (5)$$

with $M = \frac{1}{2}(\phi_{kl} + \phi_*)$ an average variable and

$$D_{KL}(\phi_{kl}||M) = \sum_{t=1}^V \phi_{klt} [\log_2 \phi_{klt} - \log_2 p(M = t)], \quad (6)$$

where V is the size of the distinct vocabulary terms and ϕ_{klt} is defined in equation 2.

VI. EXPERIMENTAL SET UP

In order to evaluate the performances of the JTV model, we utilize three types of multiple contrastive viewpoint text data:

TABLE III. STATISTICS ON THE SIX USED DATA SETS

Viewpoint	OC		AW		GM1		GM2		IP1		IP2	
	for	Ag	allow	not	illegal	not	hurt	no	pal	is	pal	is
Number of documents	434	508	213	136	44	54	149	301	149	149	148	148
Total number of tokens	14594		44482		10666		47915		209481		247059	
Average number of tokens per document	15.49		127.45		108.83		106.47		702.95		834.65	

(1) short-text data where people on average express their viewpoint briefly with few words like survey’s verbatim response or social media posts; (2) 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 six different data sets, corresponding to different contention issues. The JTV code and all data sets are publicly available⁵. We extended the Mallet toolkit⁶. Table III describes the used data sets.

ObamaCare (OC)⁷ 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 II is a human-made summary of this corpus.

Assault Weapons (AW)⁸: includes posts extracted from “debate.com”. The contention question is “Should assault weapons be allowed in the United States as means of allowing individuals to defend themselves?”. The viewpoints are either “should be allowed” or “should not be allowed”.

Gay Marriage 1 (GM1)⁹: contains posts from “debate.com” related to the contention question “Should gay marriage be illegal?”. The posts’ stance are either “should be illegal” or “should be legal”.

Gay Marriage 2 (GM2)¹⁰: 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) 1 and 2¹¹: are two datasets extracted from BitterLemons web site. Israel-Palestine 1 contains articles of two permanent editors, a Palestinian and an Israeli, about the same issues. 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”. Israel-Palestine 2 contains also weekly articles about the same issues from different Israeli and Palestinian guest authors invited by the editors to convey their views sometimes in form of interviews. Note that each issue, in these data sets’ articles, corresponds to a different contention question. Although this does not correspond to our input assumption (i.e., all documents discuss

the same contention issue), we are exploring this corpus to measure the scalability of our method for long editorial documents. Moreover, this is a well-known data set used by most of the previous related work in contention [4], [25], [18].

Paul et al. [18] stress 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 [26] 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 a different number of topics K . The number of viewpoints L is equal to 2. The number of the Gibbs Sampling iterations is 1000. The TAM model [18] (Section III) and LDA [20] are run as a means of comparison during the evaluation. TAM parameters are set to their default values with same number of topics and viewpoints as JTV. LDA is run with a number of topics equal to twice the number of JTV’s topics K , $\beta = 0.01$ and $\alpha = 50/2K$.

VII. MODEL EVALUATION

A. Qualitative Evaluation

We perform a simultaneous qualitative analysis of the generated topic-viewpoint pairs (i.e., arguing expressions) by the JTV model and their clustering (Section V) according to the viewpoint they convey. The analysis is illustrated by using the ObamaCare data set. Table IV presents an example of the result output produced by the clustering component which uses the inferred topic-viewpoint pairs as input. The number of topics and the number of viewpoints (clusters) are set to $K = 5$ and $L = 2$ respectively. Each one of these clusters is represented by a collection of topic-viewpoint pairs automatically generated and assigned to it. Each topic-viewpoint in a given cluster (e.g., Topic 1-Viewpoint 1) is represented by the set of top terms. The terms are sorted in descending order (from left to right) according to their probabilities. We try to qualitatively observe the viewpoint coherence of clustered arguing expressions as well as their intrinsic topicality coherence.

In Table IV, the majority of the topic-viewpoint pairs, corresponding to the same viewpoint, are most likely conveying the same stance and discussing topics similar to those in the ground truth summary of the corpus (Table II). For instance, taking a closer look at the original data suggests that Topic 1-Viewpoint 1 (Table IV) argues that many people, like poor or elderly people, cannot afford a healthcare or pay for it. This can correspond to the first support arguing expression in Table II. Similarly, Topic 2-Viewpoint 1 discusses the need for an

⁵<https://webdocs.cs.ualberta.ca/~atrabels/ICDM2014Code/>

⁶<http://mallet.cs.umass.edu/>

⁷<http://www.gallup.com/poll/126521/favor-oppose-obama-healthcare-plan.aspx>

⁸<http://www.debate.org/opinions/should-assault-weapons-be-allowed-in-the-umited-states-as-means-of-allowing-individuals-to-defend-themselves>

⁹<http://www.debate.org/opinions/should-gay-marriage-be-illegal>

¹⁰http://www.createdebate.com/debate/show/How_can_gay_marriage_hurt_any_one

¹¹<http://www.bitterlemons.net/>

TABLE IV. CLUSTERING OUTPUT USING THE JTV’S GENERATED TOPICS-VIEWPOINTS FROM OBAMACARE DATA SET AS INPUT.

Viewpoint 1						
Topic 1	people	cant_afford	pay	poor	elderly	healthcare
Topic 2	healthcare	people	coverage	access	years	affordable
Topic 3	insurance	health	companies	medical	public	premiums
Topic 4	healthcare	people	dont_have	uninsured	doctors	prices
Topic 5	healthcare	country	system	afford	world	children
Viewpoint 2						
Topic 1	cost	increase	expensive	reason	problem	main
Topic 2	government	control	economy	expensive	involved	private
Topic 3	bill	feel	plan	start	social	read
Topic 4	dont_think	healthcare	good	work	fair	debt
Topic 5	money	pay	medicare	dont_know	medicine	dont_want

affordable and accessible coverage for people which can be matched with the fifth support arguing expression in Table II. Topic 3-Viewpoint 1 expresses the urgency of stopping the insurance companies premiums from increasing (e.g., “*I think if they don’t do anything about healthcare, the insurance companies will continue to increase their premiums until no one can afford it*”, “*I do want the public option (...) we have to keep our insurance company at lower premiums because their prices are skyrocketing*”)¹². This may correspond to the sixth support arguing expression in Table II. Moreover, a query of the original data text using some of Topic 5-Viewpoint 1 terms suggests that the topic is about criticizing the healthcare system in the country (U.S.A) (e.g., “*Because the greatest country in the world has a dismal healthcare system*”, “*The biggest country in the world not to have healthcare for their people. I’m set but my children won’t be.*”). A match can be established with the second support arguing expression in Table II.

Like Viewpoint 1, Viewpoint 2 contains viewpoint-coherent arguing expressions. For instance, Topic 1-Viewpoint 2 emphasizes the problem of increasing costs that might be yielded by the bill which is also expressed as an opposing arguing expression in the ground-truth summary. Topic 2-Viewpoint 2 refers to government control, involvement and economy which is a vocabulary usually used by people opposing the bill (see fourth and fifth oppose arguing expressions in Table 2). Topic 4-Viewpoint 2 conveys the belief that the bill will not work or that it is not fair or good, e.g., “*I don’t think it’s going to work*”, “*I don’t think it’s fair*”, “*I don’t think it is good*”. It also opposes the bill because of the debt that it may induce e.g., “*I don’t think we can pay for it; we are going in debt to pay for it*”. Topic 5-Viewpoint 2 argues the unwillingness of people to pay for others, e.g., “*I don’t wanna have money taken out of my check to pay for people who won’t work*”, and also the problems that may be induced on medicare, e.g., “*Medicare pays hardly anything to the doctors. He[Obama]’s going to cut the medicare as I understand it, and less doctors will take it and medicare will be more limited and the healthcare costs will go up.*” Although, this topic-viewpoint seems to be coherent in term of the viewpoint it voices (oppose), its topicality coherence is questionable. One may think that the medicare problem should be included in Topic 1-Viewpoint 2. Topic 3-Viewpoint 2 may also produce some topicality ambiguity despite a coherence in the viewpoint, e.g., “*I feel like this is socialized medicine*”, “*Just a bad bill, too expensive. Nobody*

¹²The quoted phrases are excerpts from the original data text. It is important to notice that two excerpts from the same document can denote different topic-viewpoint pairs.

has read it.”. These topicality incoherences exist even when we increase the number of topics. They need to be addressed in our future work.

B. Quantitative Evaluation

We perform three tasks. In the first task we assess how well our model fits six different data sets. In the second task we evaluate how well it is able to generate distinct topic-viewpoint pairs. In the third task we appraise our model accuracy in classifying documents according to their viewpoints and hence judge the discriminative power of the model’s features in distinguishing the viewpoint of a document. For the three tasks we benchmark our model against TAM, which incorporates the topic-viewpoint dimension, as well as against the LDA model. The number of topics given as input to LDA is equal to the number of topic-viewpoint pairs. For the evaluation procedure we use three metrics.

1) *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 [23]. We compute the perplexity under estimated parameters of JTV and compare it to those of TAM and LDA for our six unigrams data sets (Section VI). Figure 2 exhibits, for each corpus, the perplexity plot as function of the number of topics K for JTV, TAM and LDA. For a proper comparison the number of topics of LDA is set to $2K$. 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 computes 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. We can also observe that the JTV’s perplexity tend to reach its minimal values for a smaller number of topics than LDA for short and medium length text. For large text, JTV and LDA perplexities are very similar.

2) *Kullback-Leibler Divergence*: Kullback-Leibler (KL) Divergence is used to measure the degree of separation between two probability distributions (see Equation 6)¹³. We utilize it for two purposes. The first purpose is to empirically

¹³Here D_{KL} is computed using the natural logarithm instead of the binary logarithm.

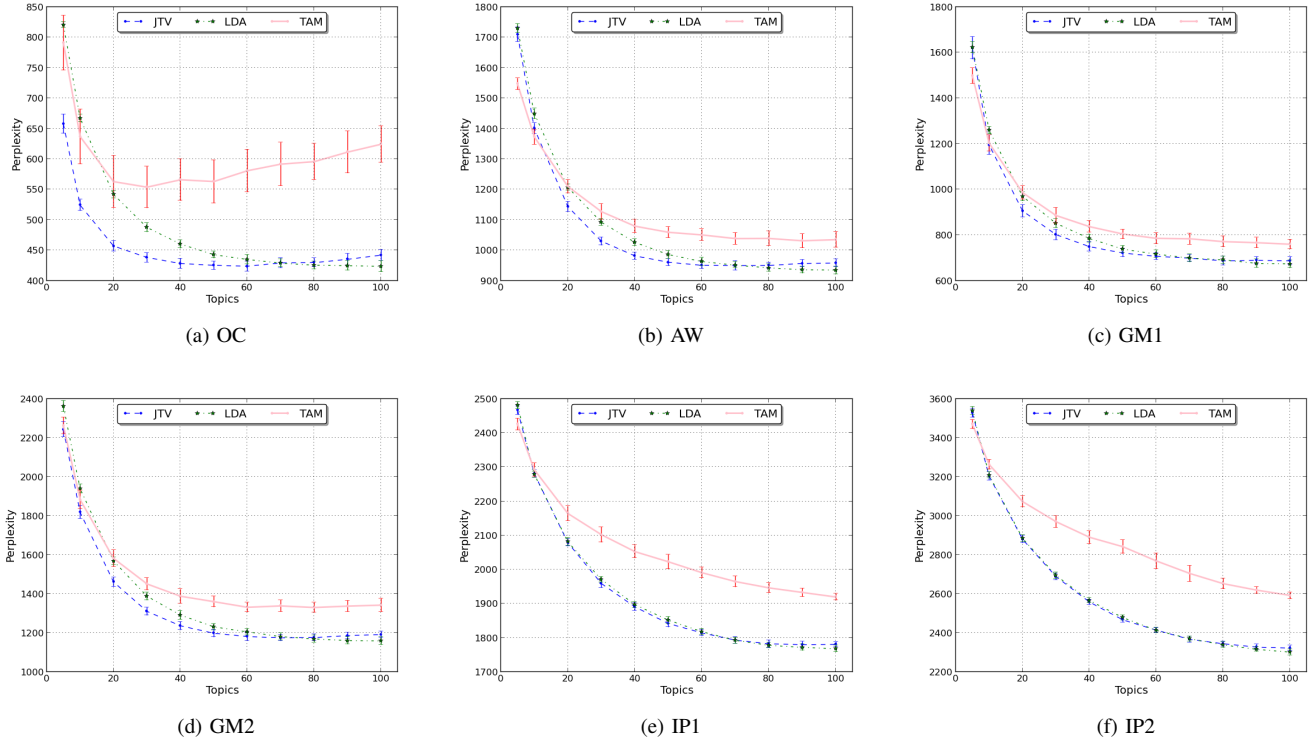


Fig. 2. JTV, LDA and TAM’s perplexity plots for six different datasets (lower is better)

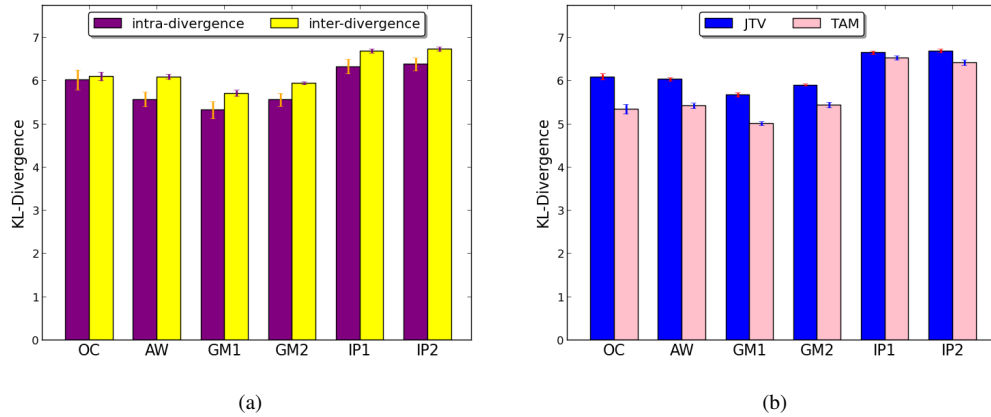


Fig. 3. Histograms of: (a) average topic-viewpoint intra/inter divergences of JTV; (b) average of overall topic-viewpoint divergences of JTV and TAM for six datasets ($K = 5$).

validate the assumption on which the clustering algorithm in Section V is based. The assumption 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 $\phi_{k'l'}$, 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 six 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 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 dimensions by JTV and TAM. This is an indicator of a good aggregation of arguing expressions. For a proper comparison, we do not assess LDA’s distinctiveness as this latter does not model the hidden viewpoint variable. 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 II).

3) *Classification Accuracy*: We take advantage of the available viewpoint labels for each document in our six datasets (Table III) in order to evaluate the quality of the generated JTV’s topic-viewpoint pairs. Recall that these topic-viewpoint dimensions are induced in a completely unsupervised manner. We adopt a classification approach where the task consists of predicting the viewpoint of a document given its learned topic-viewpoint proportions (see Section IV) as features. Topic-viewpoint proportions for each document are derived from JTV’s topic-viewpoint assignments of each word in the document. Similarly, the topic and viewpoint proportions yielded by TAM and the topic proportions induced by LDA are computed. It is important to note that classifying documents according to their viewpoints or inferring the right label in unsupervised manner is not the intent of our study. The classification is only performed as means of validation of the JTV’s modeling of the viewpoint dimension, as well as, of comparison with TAM in this regard. Indeed, the objective of the task is to assess the discriminative power of the models’ features in distinguishing the viewpoint of a document. A better discriminative power would denote a better grasping of the hidden viewpoint concept by the topic model. This evaluation procedure can also be used to check the effectiveness of the document dimensionality reduction into a topic-viewpoint space. For the classification, we used the support vector classifier in the Weka framework with the Sequential Minimal Optimization method (SMO). We compare the accuracies of the classification obtained when using JTV features (topic-viewpoint proportions), TAM features (topic proportions + viewpoint proportions) and LDA’s features (topic proportions). During this task, we perform a uniform under-sampling of the Assault Weapon (AW) and Gay Marriage 2 (GM2) datasets in order to have a balanced number of opposed viewpoint for supervision. Thus, the baseline accuracy is exactly 50% for all data sets except for the ObamaCare, 54%, and the Gay Marriage 1, 55%. We run the JTV, TAM and LDA models 20 times on all data sets and, in each run, we compute the accuracy of a 10 fold cross-validation procedure. The average accuracies for all data sets are shown in Figure 4. For each data set, the plot reports the best accuracy yield by any number of topics K as input, for each model, along with the accuracies for $K = 5$ and $K = 100$.

Although the accuracies differ from one data set to another, the best accuracies using the features generated by JTV are higher than the baselines and the best accuracies yielded by LDA or TAM features for all six datasets. Thus, JTV features (topic-viewpoint proportions) have more discriminative power to distinguish the viewpoints of contentions documents than TAM or LDA features. We also observe that most of the JTV’s peaks are reached quicker (i.e., for a smaller number of topics) than the competing models. This means that the JTV model has the capacity of accurately and efficiently reducing the contentious document space more than TAM and LDA.

VIII. CONCLUSION AND FUTURE WORK

We suggested a fine grained probabilistic framework for improving the quality of opinion mining from different types of contention texts. We proposed a Joint Topic Viewpoint model (JTV) for the unsupervised detection of arguing expressions. Unlike common approaches, the proposed model focuses on arguing expressions that are implicitly described in unstructured text according to the latent topics they discuss and the implicit viewpoints they voice. We also implemented a clustering algorithm which gets as input the learned topic-viewpoint pairs from JTV and group them according to their voiced viewpoint. The qualitative and quantitative assessments of the model’s output show a good capacity of JTV in handling different contentious issues when compared to similar models. Moreover, analysis of the experimental results shows the effectiveness of the proposed model to automatically and accurately detect recurrent and relevant patterns of arguing expressions.

JTV assumes that each topic is discussed with different proportions according to the endorsed viewpoint. Some topics may be specific to only one particular viewpoint. In this case, the corresponding generated topic-viewpoint pairs can be redundant or contain incoherent topical information. This would later mislead the arguing expression clustering task. Future work should relax this assumption in order to enhance the topicality and viewpoint coherence of extracted topic-viewpoint pairs, as well as the arguing phrases. Moreover, automatically finding the optimal numbers of topics and viewpoint remains an open problem. Extension of JTV based on Nonparametric Bayesian models, e.g., Hierarchical Dirichlet Processes [27], can be considered.

Another future study needs to focus on the generation of the snippet-summary of arguing expressions (see Table II) given the generated topic-viewpoint terms and the clustering output (see Table IV). In the qualitative evaluation (Section VII-A), the induced terms were used to search for potential relevant sentences to the topic-viewpoint in the original corpus. The task of selecting the adequate informative snippets or sentences from a query to the original text should be automated using extractive summary and information retrieval techniques. Moreover, the coherence of generated arguing expressions terms, and the summaries, needs to be assessed (human-oriented evaluation). The reference summaries as a ground truth of the used contentious corpora, or the issues themselves, have to be created by human experts for the automatic summary evaluation.

REFERENCES

- [1] J. M. Jones, “In u.s., 45% favor, 48% oppose obama healthcare plan,” Mar. 2010. [Online]. Available: <http://www.gallup.com/poll/126521/favor-oppose-obama-healthcare-plan.aspx>
- [2] F. H. van Eemeren, *Crucial Concepts in Argumentation Theory*. Amsterdam University Press, August 2001.
- [3] S.-M. Kim and E. H. Hovy, “Crystal: Analyzing predictive opinions on the web,” in *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007, pp. 1056–1064.
- [4] W.-H. Lin, T. Wilson, J. Wiebe, and A. Hauptmann, “Which side are you on?: Identifying perspectives at the document and sentence levels,” in *Proceedings of the Tenth Conference on Computational Natural Language Learning*, 2006, pp. 109–116.

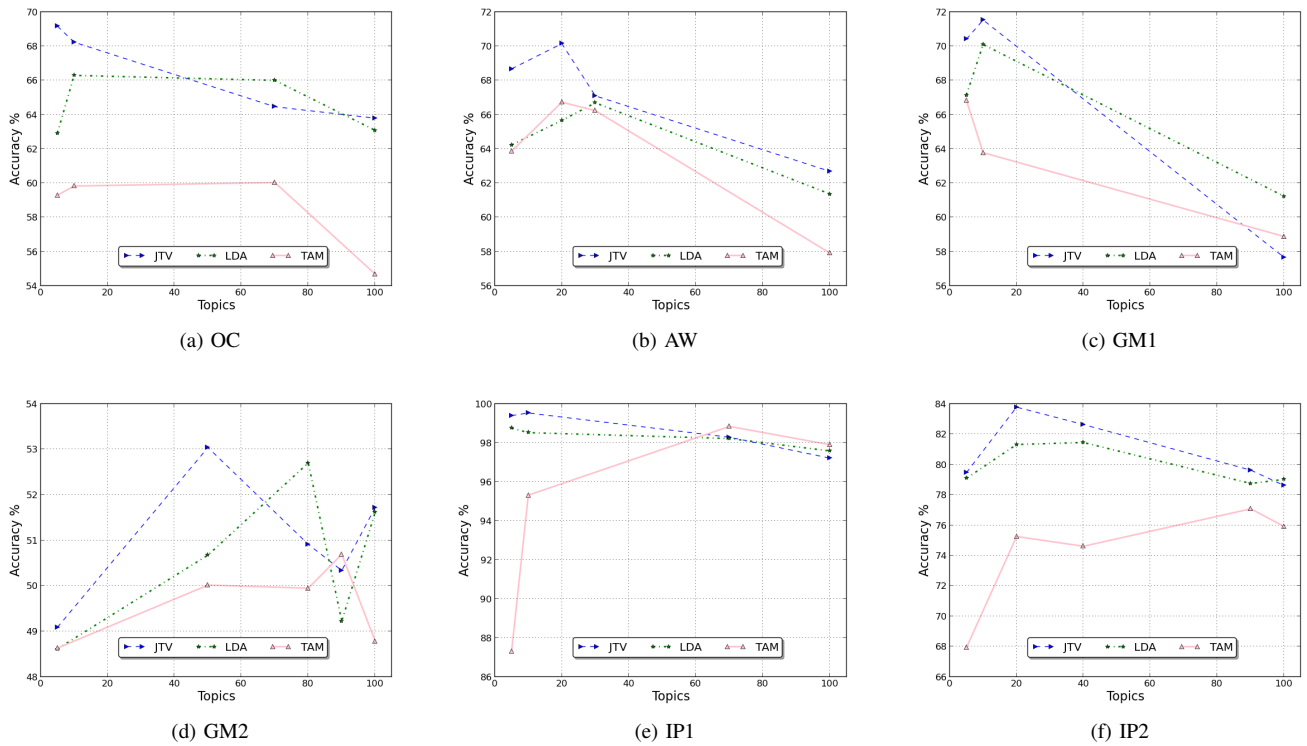


Fig. 4. JTV, LDA and TAM’s features classification accuracies plots for six different data sets

- [5] S. Park, K. Lee, and J. Song, “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*, 2011, pp. 340–349.
- [6] M. Thomas, B. Pang, and L. Lee, “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*, 2006, pp. 327–335.
- [7] S. Somasundaran and J. Wiebe, “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*, 2010, pp. 116–124.
- [8] Y. Jo and A. H. Oh, “Aspect and sentiment unification model for online review analysis,” in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, 2011, pp. 815–824.
- [9] I. Titov and R. McDonald, “Modeling online reviews with multi-grain topic models,” in *Proceedings of the 17th International Conference on World Wide Web*, 2008, pp. 111–120.
- [10] W. X. Zhao, J. Jiang, H. Yan, and X. Li, “Jointly modeling aspects and opinions with a maxent-lda hybrid,” in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010, pp. 56–65.
- [11] C. Lin and Y. He, “Joint sentiment/topic model for sentiment analysis,” in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 2009, pp. 375–384.
- [12] W.-H. Lin, E. Xing, and A. Hauptmann, “A joint topic and perspective model for ideological discourse,” in *Machine Learning and Knowledge Discovery in Databases*, W. Daelemans, B. Goethals, and K. Morik, Eds. Springer Berlin Heidelberg, 2008, vol. 5212, pp. 17–32.
- [13] A. Mukherjee and B. Liu, “Mining contentions from discussions and debates,” in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 841–849.
- [14] —, “Discovering user interactions in ideological discussions,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 2013, pp. 671–681.
- [15] M. Qiu and J. Jiang, “A latent variable model for viewpoint discovery from threaded forum posts,” in *Proceedings of NAACL-HLT*, 2013, pp. 1031–1040.
- [16] Y. Fang, L. Si, N. Somasundaram, and Z. Yu, “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*, 2012, pp. 63–72.
- [17] S. Gottipati, M. Qiu, Y. Sim, J. Jiang, and N. A. Smith, “Learning topics and positions from debatepedia,” in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2013.
- [18] M. J. Paul, C. Zhai, and R. Girju, “Summarizing contrastive viewpoints in opinionated text,” in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010, pp. 66–76.
- [19] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
- [20] T. L. Griffiths and M. Steyvers, “Finding scientific topics,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 1, pp. 5228–5235, 2004.
- [21] S. Basu, I. Davidson, and K. Wagstaff, *Constrained Clustering: Advances in Algorithms, Theory, and Applications*, 1st ed. Chapman & Hall/CRC, 2008.
- [22] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, “Constrained k-means clustering with background knowledge,” in *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, pp. 577–584.
- [23] G. Heinrich, “Parameter estimation for text analysis,” Fraunhofer IGD, Tech. Rep., September 2009.
- [24] S. Kullback and R. A. Leibler, “On information and sufficiency,” *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.
- [25] M. J. Paul and R. Girju, “A two-dimensional topic-aspect model for discovering multi-faceted topics,” in *Proceedings of AAAI*, 2010.
- [26] M. Steyvers and T. Griffiths, “Probabilistic topic models,” *Handbook of Latent Semantic Analysis*, vol. 427, no. 7, pp. 424–440, 2007.
- [27] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, “Hierarchical dirichlet processes,” *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.