Knowl Inf Syst (2016) 48:537–560
DOI 10.1007/s10115-015-0888-6

REGULAR PAPER

Mining contentious documents

Amine Trabelsi[1] · **Osmar R. Zaïane**[1]

Received: 2 March 2015 / Revised: 22 June 2015 / Accepted: 5 October 2015 /
Published online: 17 October 2015

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 analyze the underlying divergent arguing expressions that may be present in a collection of contentious documents. It extends the original Latent Dirichlet Allocation, 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 analyses 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. Furthermore, the coherence of these arguing expressions is proved to be of a high quality when evaluated on the basis of recently introduced automatic coherence measure.

Keywords Arguing expressions detection · Contentious text analysis · Unsupervised clustering · Opinion mining · Automatic coherence measure for topic models

1 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

✉ Amine Trabelsi
atrabels@ualberta.ca

[1] Department of Computing Science, University of Alberta, Edmonton, Canada

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

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

reviews. In addition to sentiment expressed toward products, other online text sources such as opinion polls, debate Web sites 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 Web sites and editorials. Mining and summarizing these resources are 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 Dictionaries).[1] Documents such as survey reports, debate site 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.[2]

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 1, 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 kinds of arguing expressions, according to the topics they invoke and the viewpoints they convey, are 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 about people's opinion regarding an issue

[1] http://www.oxfordlearnersdictionaries.com/definition/english/contentious.

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

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 2 states the problem. Section 3 explains the key issues in the context of recent related work. Section 4 provides the technical details of the proposed model, the Joint Topic Viewpoint model (JTV). Section 5 describes the clustering task used to obtain a feasible solution. Section 6 provides a description of the experimental setup 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 discusses the future work. A shorter version of this paper was originally published in [34].

2 Problem statement

This paper examines the task of mining the underlying topics and the hidden viewpoints of arguing expressions toward the summarization of contentious text. An example of a human-made summary of arguing expressions [14] on, what is commonly known as the Obama healthcare reform, is presented in Table 2. 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 2 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 health care. 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 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

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 health care more affordable	Government should not be involved in healthcare
Don't trust insurance companies	Cost the government too much

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 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 $\{\text{doc}_1, \text{doc}_2, \ldots, \text{doc}_D\}$, where each document doc_d expresses one or more viewpoints $\vec{v}^d$ from a set of L possible viewpoints $\{v_1, v_2, \ldots, v_L\}$, 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 $\{\phi_{1l}, \phi_{2l}, \ldots, \phi_{Kl}\}$, 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, $k = 1 \ldots K$, 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."

3 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 [35]. 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 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 microblogs. 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 [15,18]. 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 [23,32]. 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 [29]. Somasundaran and Wiebe [29] classify ideological stances in online debates using generated arguing clues from the Multi-Perspective Question Answering (MPQA) opinion corpus.[3] 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 [13,33,37]. Lin and He [17] 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. [19] 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 [21,22] examine mining contention from discussion forums data where the interaction between different authors is pivotal. They attempt to discover agreement/disagreement (or contention/agreement) indicators called AD (or CA) expressions using three different Joint Topic Expressions models (JTE). Examples of agreement expressions are "I agree," "rightly said," "very well put" or "I do support." Examples of disagreement expressions are "I contest," "I really doubt," "Can you prove" or "you have no clue." The found expressions are excerpts from conversational text data, mainly debate forums. In debate forums, the posts' authors are often referring, citing and responding to each

[3] http://mpqa.cs.pitt.edu/.

other. The proposed versions of JTE [21,22] model the author pairs discussing a contention in order to be able to classify the nature of interaction in a post-topic modeling stage. However, these proposals do not model the authors' viewpoint dimension, unlike JTV. For JTE, the objective is not to summarize the main reasons held by authors of divergent viewpoints. The goal is to find the lexicon that people usually use to express agreement or disagreement. Detected agreement and disagreement expressions by JTE are used to discover points of contention but not to separately summarize divergent viewpoints on a particular topic of contention. Moreover, JTE versions are very dependent of a supervised component, the Maximum Entropy model. It helps in initializing the detection of AD expressions.

Qiu and Jiang [26] 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 toward 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. [8] 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, it cannot cluster documents according to their viewpoints. Moreover, the discovered topics are not necessarily of contention. Recently, Gottipati et al. [9] propose a topic model to infer human interpretable text in the domain of issues using Debatepedia[4] 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. [25]. It introduces the problem of contrastive summarization which is very similar to our stated problem in Sect. 2 and proposes the Topic Aspect Model (TAM). Throughout the experiments that we present in the following sections, we will often use TAM as a conventional comparison method to JTV. The contrastive summarization consists of summarizing the contentious text by detecting the relevant sentences describing each of the possible expressed viewpoint. TAM models this viewpoint as a topic model's variable, like JTV, which leads to topic–viewpoint dimensions as output, i.e., a list of word with both a topic and viewpoint assignments. Therefore, this provides a valuable framework for comparison with JTV. Moreover, TAM and JTV are mainly unsupervised methods, which enables a fair comparison.

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

Paul et al. [25] use the output distributions of TAM 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

The goal of most conventional clustering and modeling approaches of text corpora is to find short descriptions and reduce the original text into its most important words and their statistical relationships. A notable approach in that regard is the Latent Semantic Indexing or Analysis (LSI) [5]. LSI is based on a linear algebra dimensionality reduction method, the Singular Value Decomposition (SVD). It takes an $N \times D$ matrix of weights of N words in D documents. The weights are usually *tf-ifd* measures [28]. It returns three matrices interpreted as the weights of the N words for K topics ($N \times K$ matrix), the weight of K topics in the input ($K \times K$ diagonal matrix) and the weights of the K topics in each document ($K \times D$ matrix). LSI is a non-generative approach which may lead to the over-fitting of the input text collections.

Similar to LSI, other linear algebra methods like matrix factorization approaches have been used in data clustering [7]. For instance, Non-negative Matrix Factorization (NMF) method has been experimented on documents collection [6]. NMF is very similar to the *probabilistic* LSI (*p*LSI) [12], a stochastic alternative to LSI. NMF and *p*LSI are different algorithms which optimize the same optimization function [7]. Ding et al. [7] argue that NMF with I-divergence and *p*LSI is equivalent. However, a major limitation of matrix factorization approaches is the static modeling, which disregards the generation context of the data [20].

The *p*LSI provides a generative probabilistic model at the word level. It models a word in a document as a mixture model, where the mixture components are multinomial random variables representing topics. However, *p*LSI does not provide a generative probabilistic model at the document level [3]. Indeed, the mixing proportions are dependent of the indexes of the documents. This leads to a number of parameters of the model that grows linearly with the corpus size which may lead to over-fitting. Similarly, the model would only learn the topic mixtures for the training documents, which also makes the generalization to unseen data difficult.

Latent Dirichlet Allocation (LDA) [3] is one of the most popular probabilistic generative models used to mine large text data sets. The LDA considers the topic mixture parameters as random variables rather than a list of parameters depending on the document index. This enables to overcome the over-fitting and to better generalize on unseen documents [3]. Therefore, LDA-based model provides a more complete generative probabilistic model than *p*LSI. It takes into account the boundaries of a document when generating topics and it leads to a reduced and scalable representation. 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 Sect. 2, each one of these topics can be shared by divergent arguing expressions

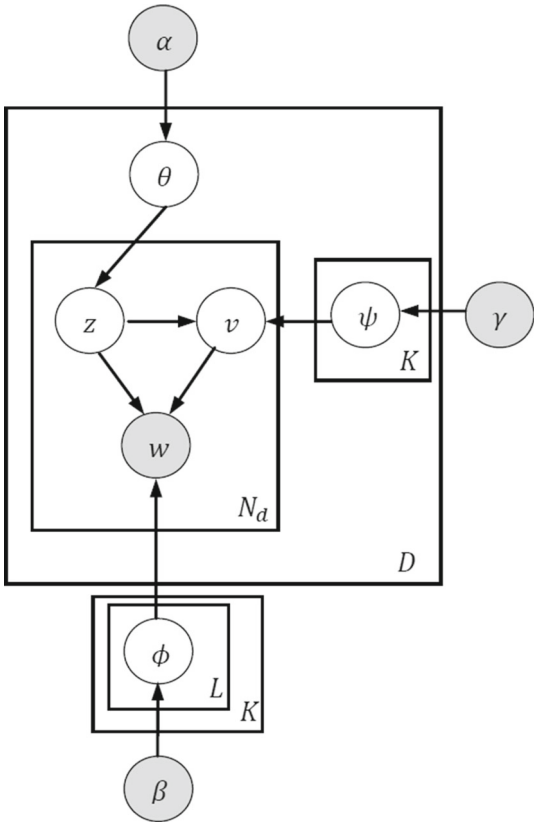

Fig. 1 The JTV's graphical model (plate notation)

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 Sect. 2, 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) 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; and 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 [3]. We use the collapsed Gibbs sampling [10], 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 $\vec{v}_{\neg i}$, $\vec{z}_{\neg i}$ and observed $\vec{w}_{\neg i}$, where $\vec{v} = \{v_i, \vec{v}_{\neg i}\}$, $\vec{z} = \{z_i, \vec{z}_{\neg i}\}$ and $\vec{w} = \{w_i, \vec{w}_{\neg i}\}$. The derived sampling equation is:

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

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:

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

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

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

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

$$\theta_{dk} = \frac{n^{(k)}_{d} + \alpha}{\sum_{k=1}^{K} n^{(k)}_{d} + K\alpha}. \tag{4}$$

5 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 Sect. 7.2.2). Therefore, we can formulate the problem of clustering arguing expressions as a constrained clustering problem [2]. The goal is to group the similar topics–viewpoints ϕ_{kl}s into L clusters (number of viewpoints), given the constraint that the L ϕ_{kl}s of the same topic k should not belong to the same cluster (cannot-link constraints). Thus, each cluster C_i where $i = 1 \ldots L$ will contain exactly K topics–viewpoints.

We suggest a slightly modified version of the constrained k-means clustering (COP-KMEANS) [36]. 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 $k^{\dagger}$ 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_{kl}s of the same topic k (lines 2–18) to be in a one-to-one matching with the set C of L clusters (lines 5–18). 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 10–16). The distance between a topic–viewpoint distribution ϕ_{kl} and another distribution ϕ_* is measured using the symmetric Jensen–Shannon distance (D_{JS}) [11] which is based on the Kullback–Leibler divergence (D_{KL}) [16]:

Algorithm 1 Topic-Viewpoint Clustering

Require: JTV's output:topic-viewpoint distributions ϕ_{kl}s, number of topics K, number of viewpoints L
1: Initialize the set C with a set of empty clusters; Choose the topic-viewpoint distributions $\phi_{k^{\dagger}1} \ldots \phi_{k^{\dagger}L}$ of the most frequent topic $k^{\dagger}$ according to JTV as the initial cluster centers.
2: **for** each topic k ($k = 1 \ldots K$) **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

Table 3 Statistics on the six used data sets

	OC		AW		GM1		GM2		IP1		IP2	
Viewpoint	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	14,594		44,482		10,666		47,915		209,481		247,059	
Average number of tokens per document	15.49		127.45		108.83		106.47		702.95		834.65	

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

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

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

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

6 Experimental setup

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 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; and (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 available.[5] We extended the Mallet toolkit.[6] Table 3 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 2 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."

[5] http://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-united-states-as-means-of-allowing-individuals-to-defend-themselves.

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 2:[11] are two data sets extracted from Bitter-Lemons 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 [18,24,25].

Paul et al. [25] stress the importance of negation features in detecting contrastive viewpoints. Thus, we performed a simple treatment of merging any negation indicators, such as "nothing," "no one" and "never," 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 [30] 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 [25] (Sect. 3) and LDA [10] 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 $\alpha = 50/2K$.

7 Model evaluation

7.1 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 (Sect. 5) according to the viewpoint they convey. The analysis is illustrated by using the Obamacare data set. Table 4 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

[9] http://www.debate.org/opinions/should-gay-marriage-be-illegal.

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

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

Table 4 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	Health care
Topic 2	Health care	People	Coverage	Access	Years	Affordable
Topic 3	Insurance	Health	Companies	Medical	Public	Premiums
Topic 4	Health care	People	Dont_have	Uninsured	Doctors	Prices
Topic 5	Health care	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	Health care	Good	Work	Fair	Debt
Topic 5	Money	Pay	Medicare	Dont_know	Medicine	Dont_want

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 Sect. 7.3, we proceed to an automatic evaluation of the coherence of our model when used with the six data sets.

In Table 4, 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 2). For instance, taking a closer look at the original data suggests that Topic 1-Viewpoint 1 (Table 4) argues that many people, like poor or elderly people, cannot afford a health care or pay for it. This can correspond to the first support arguing expression in Table 2. Similarly, Topic 2-Viewpoint 1 discusses the need for an affordable and accessible coverage for people which can be matched with the fifth support arguing expression in Table 2. 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 2. 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 (USA) (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 2.

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,"

[12] The 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.

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

7.2 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.

7.2.1 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 [11]. We compute the perplexity under estimated parameters of JTV and compare it to those of TAM and LDA for our six unigrams data sets (Sect. 6). 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 onefold perplexity from a tenfold 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.

7.2.2 Kullback–Leibler divergence

Kullback–Leibler (KL) divergence is used to measure the degree of separation between two probability distributions (see Eq. 6).[13] We utilize it for two purposes. The first purpose is to

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

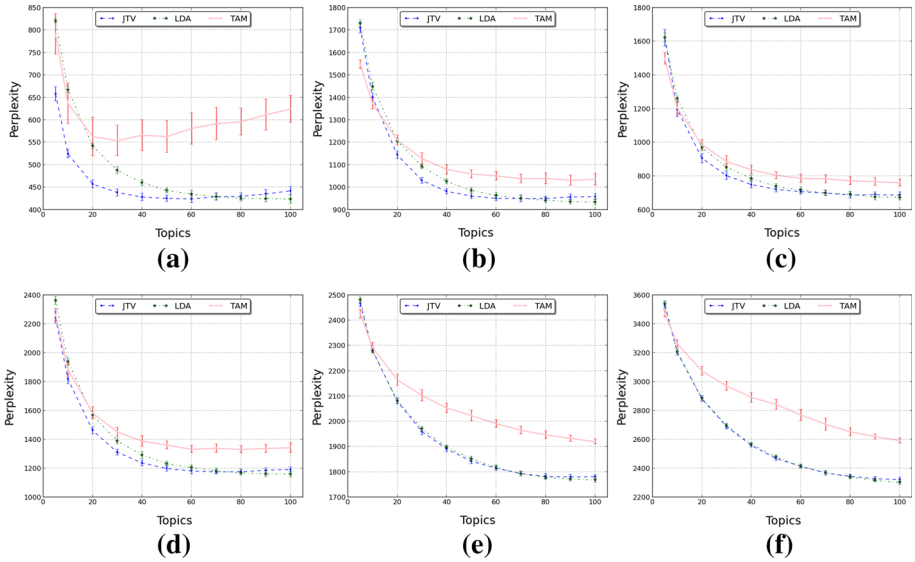

Fig. 2 JTV, LDA and TAM's perplexity plots for six different datasets (lower is better). **a** OC. **b** AW. **c** GM1. **d** GM2. **e** IP1. **f** IP2

empirically validate the assumption on which the clustering algorithm in Sect. 5 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 $\phi_{k'*}$, 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 data sets. 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 Sect. 2).

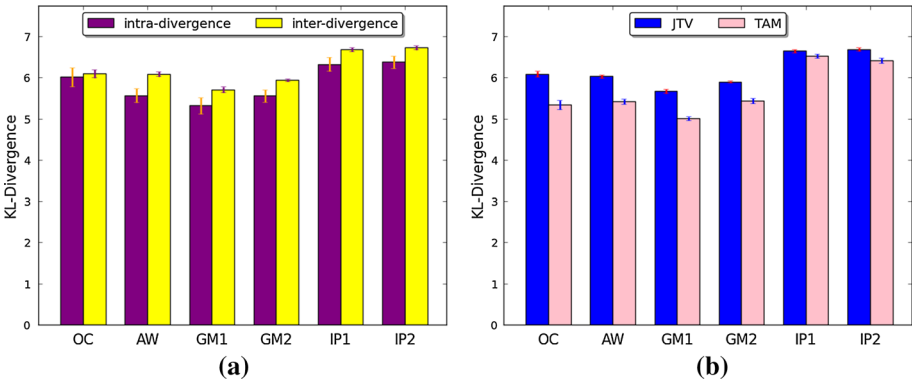

Fig. 3 Histograms of **a** average topic–viewpoint intra-/inter-divergences of JTV; **b** average of overall topic–viewpoint divergences of JTV and TAM for six datasets ($K = 5$)

7.2.3 Classification accuracy

We take advantage of the available viewpoint labels for each document in our six datasets (Table 3) 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 Sect. 4) 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 tenfold cross-validation procedure. The average accuracies for all data sets are shown in Fig. 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

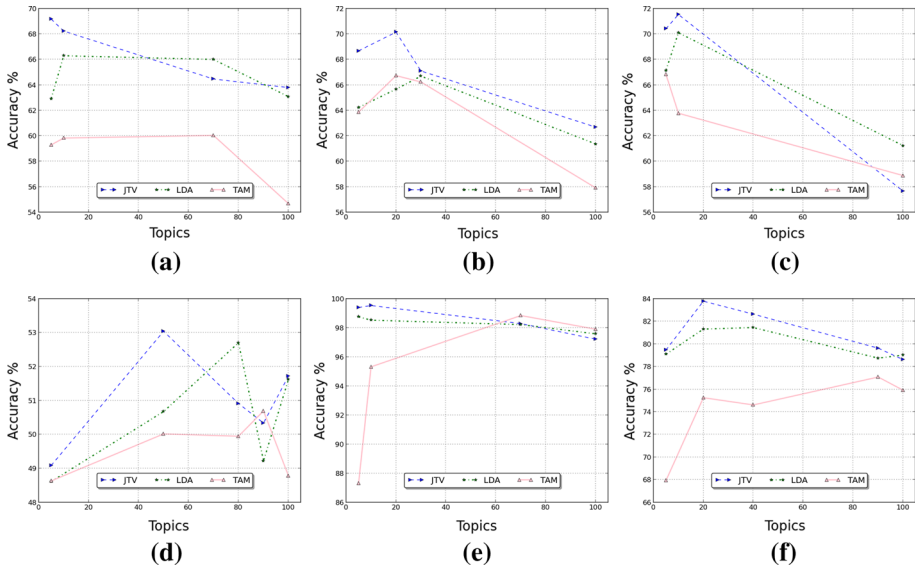

Fig. 4 JTV, LDA and TAM's features classification accuracies plots for six different data sets. **a** OC. **b** AW. **c** GM1. **d** GM2. **e** IP1. **f** IP2

LDA or TAM features for all six data sets. 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.

7.3 Automatic coherence evaluation

In Sect. 7.1, we proceed to an informal analysis of the coherence of the JTV topic–viewpoint pairs (or arguing expressions), as well as, their viewpoint coherence after the application of Algorithm 1. The Obamacare (OC) dataset is used as a case study. In this section, we automatically measure the two types of coherences for all datasets, and compare the results of the JTV to those of TAM in that respect. The first coherence is an overall assessment of all generated topic–viewpoint pairs as lists of words. The second coherence evaluates the coherence of a group of topic–viewpoint pairs with respect to a particular viewpoint. We exploit a recent work on automatic evaluation of topics coherence [27]. Indeed, Röder et al. [27] propose a unifying framework of coherence measures, Palmetto, which encompasses existing measures in the literature as well as unexplored ones.

7.3.1 Palmetto framework

The Palmetto framework[14] [27] is constituted of four separate parts or dimensions that are exchangeable and which span the configuration space of coherence measures. The input is a

[14] http://aksw.org/Projects/Palmetto.html.

set of words W which corresponds to the topic to evaluate. The output is a coherence score for the set W.

Segmentation The first part is segmentation. A coherence should measure how well pairs of single words or subsets of them are fitting together. A set W can be segmented into a set of pairs of subset of different sizes. The space of possible segmentations is denoted by S. The coherence measure computes the degree of support that the second part of the pair provides to the first part or subset. When the components of the pair are single words, the segmentation is denoted by $S^{\text{one}}_{\text{one}}$. When the first part is a single word and the second part is the exact set W, the segmentation is denoted as $S^{\text{one}}_{\text{set}}$.

Probability estimation The second dimension is the set of methods P used to estimate word probabilities given an underlying data source. The probability of a word can be estimated as the number of documents in which the word occurs divided by the number of all documents. When the number of documents is substituted with the number of sliding windows of size n, this estimated probability is called the boolean sliding window probability denoted by $P_{sw(n)}$. The sliding window method captures proximity between word tokens [27].

Confirmation measure A confirmation measure takes as input a probability method and a segmented pair $S_i = (W', W^*)$, where W' and W^* are subsets of the initial set of words W. It computes how well W^* supports W. Two main approaches are considered.

The first is the direct confirmation measure. It computes the similarity between W' and W^*. For instance, a very popular direct similarity is the Pointwise Mutual Information (PMI), called also log ratio measure $m_{\text{lr}}(W', W^*) = \log \frac{P(W', W^*)+\epsilon}{P(W^*)P(W')}$. Another example is the normalized PMI (NPMI) $m_{\text{nlr}}(W', W^*) = \frac{m_{\text{lr}}(W', W^*)}{-\log(P(W', W^*)+\epsilon}$.

The second approach is the indirect confirmation measure. It computes the similarity of words in W' and W^* with respect to direct confirmations to all words of W. Thus, W' and W^* are represented with vectors of size $|W|$ where each element is a direct confirmation measure between W' or W^* and a single word from W. These vectors are called context vector. The indirect confirmation measure is a vector similarity score. A very common measure is the cosine similarity m_{cos}, used also by Aletras and Stevenson [1] in that context. An indirect confirmation measure is denoted by $\tilde{m}$.

Aggregation The fourth dimension is the aggregation score of the confirmation measures of all segmented pairs given a particular segmentation. Examples of aggregations are the arithmetic mean σ_{a}, the median σ_{m} and the geometric mean σ_{g}.

7.3.2 Used coherence measure

The main state-of-the-art automatic coherence measures of topic models output can be modeled using the described framework. For instance, the proposed coherence by [4] which is based on the NPMI can be formulated as $C_{\text{NPMI}} = (S^{\text{one}}_{\text{one}}, P_{sw(10)}, m_{\text{nlr}}, \sigma_{\text{a}})$. For this confirmation measure, the segmented pairs are composed of single words. The probability estimation method is the boolean sliding window. The confirmation measure is the direct normalized PMI, and the aggregation score is the arithmetic mean of all pairs score.

In their experiments, Röder et al. [27] compute different existing and unexplored automated coherences of several topics (set of words). These topics are generated from previous studies on coherence evaluation. They are extracted from large data sets. A dataset includes

a corpus, a set of topics and the human ratings of those topics with respect to their interpretability and understandability. Thus, they compute a Pearson correlation between the topics rankings generated by automatic coherences scores, and rankings induced by the human ratings. A good coherence measure highly correlates with human ratings. Two different types of data sources are used in order to derive word counts and probabilities needed for automatic coherence computation. These comprise the original corpora used for topics learning and the external Wikipedia corpus. The coherence measure that correlates the most with human ratings of topics from different datasets, while using different data sources in probabilities computation, is the C_V measure. $C_V = (S^{\text{one}}_{\text{set}}, P_{sw(110)}, \tilde{m}_{\cos(\text{nlr})}, \sigma_{\text{a}})$ is an unexplored new combination measure. It combines the segmentation $S^{\text{one}}_{\text{set}}$, the boolean sliding window, the indirect cosine measure with NPMI and the arithmetic mean for aggregation.

We exploit the described framework and the main results found in Röder et al. [27] work. We adopt the C_V measure in our setting to evaluate the individual topicality coherence of our topic–viewpoint pairs through the combination of our JTV model and the constrained clustering algorithm (Algorithm 1). An example of the constrained algorithm's output is presented in Table 4. We further assess the viewpoint coherence of the formed clusters.

7.3.3 Experiments

Using the C_V measure, we evaluate the coherence of the JTV's and TAM's learned topic–viewpoint pairs from the six datasets described in Table 3. The same datasets are used as data sources to compute word counts and probabilities required for C_V computation. Experiments with Wikipedia corpus as a data source for probability estimation give lower coherence measure than those obtained with training data sets. This finding contradicts the results established by Röder et al. [27]. It could be imputed to the difference, in the nature and the structure, between the informal and specific discussion forum or contentious documents and the organized, general and formal articles contained in Wikipedia.

For probabilities estimation, we consider the six contentious datasets as data sources. In their study, Röder et al. [27] use a boolean sliding window size of 110 for C_V coherence metric. The use of the same window size in estimating the words probabilities with the training datasets as data sources results in uninterpretable coherence scores (almost all correlations to human ratings equal to 1). Moreover, a window size of 110 would not be appropriate for short documents with particularly small average length that is close to 110 tokens per document (see Table 3). Thus different window sizes of 1, 5, 10, 20 were tried out in our experiments. Size 1 gives the largest average C_V coherence value for topic–viewpoints pairs for both JTV and TAM.

The coherence of topic–viewpoint pairs generated by a particular model for a given dataset are obtained by averaging the C_V scores associated with each topic–viewpoint pair. Table 5 summarizes the results of the overall coherence for the JTV and the TAM models. The coherence scores are computed with the following parameter settings: the number of topics $K = 5$; the number of viewpoints $L = 2$; the number of top words = 10; and the number of runs = 100.

Table 5 shows that the best average coherence scores for topic–viewpoint pairs are achieved by our JTV model compared to the TAM model. The large values of the coherence scores, achieved by our JTV model, confirm the quality of the topic–viewpoint pairs that it is able to generate for different data sets.

We proceed to another experiment in order to assess the coherence of the topic–viewpoint groupings according to the constrained clustering algorithm (Algorithm 1) when the number of viewpoints is equal to 2. The idea consists of checking whether the majority of topics–

Table 5 Average coherence measures of topic–viewpoints generated by JTV and TAM models with their standard deviations

	JTV		TAM	
	Average	SD	Average	SD
OC	0.680	0.021	0.317	0.044
AW	0.897	0.010	0.648	0.045
GM1	0.587	0.021	0.283	0.061
GM2	0.882	0.010	0.690	0.041
IP1	0.903	0.003	0.851	0.006
IP2	0.888	0.004	0.841	0.006

viewpoints in one cluster are more coherent with the documents of a particular stance, while the majority of the topics–viewpoints in the second cluster happens to be more coherent with the documents of opposing stance. The divergence, in that case, is an indicator of a good viewpoint grouping.

Algorithm 2 Checking the divergence of learned viewpoints

Require: Coherence measure C_V, Corpus $D1$ of documents labeled as stance1, Corpus $D2$ of documents labeled as stance2, Learned topics–viewpoints (t-v)s of a model, A number of viewpoints L equal to 2.

for each viewpoint v_j in v_1, v_2 **do**
  for each topic-viewpointtopic-viewpoint t_i-v_j **do**
    compute $C_{V1} = C_V(t_i\text{-}v_j)$, s.t. $D1$ is used for probability estimation
    compute $C_{V2} = C_V(t_i\text{-}v_j)$, s.t. $D2$ is used for probability estimation.
    if $C_{V1} > C_{V2}$ **then**
      label t_i-v_j with 1
    else
      if $C_{V1} < C_{V2}$ **then**
        label t_i-v_j with 2
      else
        label t_i-v_j with Random(1,2)
      end if
    end if
  end for
  if the majority of t_i-v_j labels is 1 **then**
    label v_j with 1
  else
    if the majority of t_i-v_j labels is 2 **then**
      label v_j with 2
    else
      label v_j with Random(1,2)
    end if
  end if
end for
if v_1 and v_2 labels are different **then**
  return True
else
  return False
end if

Algorithm 2 explains in detail how to check this type of divergence given the topic–viewpoint pairs generated by a model, the coherence measure C_V and two corpora $D1$ and

Table 6 Viewpoint divergence rates derived after 100 runs of Algorithm 2

	JTV (%)	TAM (%)
OC	75	41
AW	76	25
GM1	67	14
GM2	43	31
IP1	82	66
IP2	52	66

$D2$ of opposed stances. For each topic–viewpoint, the algorithm computes two C_V scores C_{V1} and C_{V2}. These coherence measures are computed by using word probabilities obtained from data sources $D1$ and $D2$, respectively. Then, each topic–viewpoint is labeled with the stance of the corpus that gives the largest coherence measures. The group of topic–viewpoint pairs sharing the same viewpoint is labeled according to the majority label of its composing elements. When the two possible groups are labeled differently, the algorithm returns a boolean true value for divergence, otherwise false.

We run the algorithm several times to determine the divergence rate of the clustered groups of topics–viewpoints. The algorithm is run with both the combination of JTV and Algorithm 1 (JTV + constrained clustering) and TAM for purpose of comparison. Table 6 reports the rates of divergence after 100 runs of the combined JTV + constrained clustering and TAM models. Our combination outperforms TAM with respect to five datasets (OC, AW, GM1, GM2 and IP1). The differences in divergence rates, in this case, are significant, reaching an average of 33 %. For the Israel–Palestine 2 dataset, TAM seems to achieve a slightly better performance. In fact, the structure of documents contained in this corpus is different from the one corresponding to the documents in the remaining five corpora. It mostly includes interview articles in the form of question–answer pairs. This may explain the obtained low rate of viewpoint divergence in the case of JTV + constrained clustering combination.

8 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. The automatic coherence evaluation, using the newly introduced framework [27], demonstrates a decent interpretability of arguing expressions generated by the combination JTV and the constrained clustering algorithm.

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. Extention of JTV based on nonparametric Bayesian models, e.g., hierarchical Dirichlet processes [31], can be considered.

Another future study needs to focus on the generation of the snippet–summary of arguing expressions (see Table 2) given the generated topic–viewpoint terms and the clustering output (see Table 4). In the qualitative evaluation (Sect. 7.1), 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 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. Aletras N, Stevenson M (2013) Evaluating topic coherence using distributional semantics. In: Proceedings of the 10th international conference on computational semantics, pp 13–22
2. Basu S, Davidson I, Wagstaff K (2008) Constrained clustering: advances in algorithms, theory, and applications, 1st edn. Chapman & Hall/CRC, London
3. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
4. Bouma G (2009) Normalized (pointwise) mutual information in collocation extraction. In: Proceedings of the GSCL, pp 31–40
5. Deerwester SC, Dumais ST, Landauer TK, Furnas GW, Harshman RA (1990) Indexing by latent semantic analysis. JASIS 41(6):391–407
6. Ding C, He X, Simon HD (2005) On the equivalence of nonnegative matrix factorization and spectral clustering. In: Proceedings of the SIAM data mining conference, pp 606–610
7. Ding C, Li T, Peng W (2008) On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. Comput Stat Data Anal 52(8):3913–3927
8. Fang Y, Si L, Somasundaram N, Yu Z (2012) Mining contrastive opinions on political texts using cross-perspective topic model. In: Proceedings of the fifth ACM international conference on Web search and data mining, pp 63–72
9. Gottipati S, Qiu M, Sim Y, Jiang J, Smith NA (2013) Learning topics and positions from debatepedia. In: Proceedings of conference on empirical methods in natural language processing, pp 1858–1868
10. Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci USA 101(1):5228–5235
11. Heinrich G (2009) Parameter estimation for text analysis. Technical report, Fraunhofer IGD
12. Hofmann T (1999) Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, pp 50–57
13. Jo Y, Oh AH (2011) 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
14. Jones JM (2010) In US, 45% favor, 48% oppose obama healthcare plan. http://www.gallup.com/poll/126521/favor-oppose-obama-healthcare-plan.aspx
15. Kim SM, Hovy EH (2007) 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
16. Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22(1):79–86
17. Lin C, He Y (2009) Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM conference on information and knowledge management, pp 375–384
18. Lin WH, Wilson T, Wiebe J, Hauptmann A (2006) 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
19. Lin WH, Xing E, Hauptmann A (2008) A joint topic and perspective model for ideological discourse. In: Daelemans W, Goethals B, Morik K (eds) Mach Learn Knowl Discov Databases, vol 5212. Springer, Berlin Heidelberg, pp 17–32

20. Mackey LW, Weiss D, Jordan MI (2010) Mixed membership matrix factorization. In: Proceedings of the 27th international conference on machine learning, pp 711–718
21. Mukherjee A, Liu B (2012) Mining contentions from discussions and debates. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, pp 841–849
22. Mukherjee A, Liu B (2013) Discovering user interactions in ideological discussions. In: Proceedings of the 51st annual meeting of the association for computational linguistics, pp 671–681
23. Park S, Lee K, Song J (2011) 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
24. Paul MJ, Girju R (2010) A two-dimensional topic-aspect model for discovering multi-faceted topics. In: Proceedings of the AAAI conference on artificial intelligence, pp 545–550
25. Paul MJ, Zhai C, Girju R (2010) Summarizing contrastive viewpoints in opinionated text. In: Proceedings of the conference on empirical methods in natural language processing, pp 66–76
26. Qiu M, Jiang J (2013) A latent variable model for viewpoint discovery from threaded forum posts. In: Proceedings of NAACL: North American chapter of the ACL-human language technologies, pp 1031–1040
27. Röder M, Both A, Hinneburg A (2015) Exploring the space of topic coherence measures. In: Proceedings of the eighth ACM international conference on Web search and data mining, pp 399–408
28. Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. Inf Process Manag 24(5):513–523
29. Somasundaran S, Wiebe J (2010) Recognizing stances in ideological on-line debates. In: Proceedings of the NAACL-HLT workshop on computational approaches to analysis and generation of emotion in text, pp 116–124
30. Steyvers M, Griffiths T (2007) Probabilistic topic models. Handb Latent Semant Anal 427(7):424–440
31. Teh YW, Jordan MI, Beal MJ, Blei DM (2006) Hierarchical dirichlet processes. J Am Stat Assoc 101(476):1566–1581
32. Thomas M, Pang B, Lee L (2006) Get out the vote: determining support or opposition from congressional floor-debate transcripts. In: Proceedings of the conference on empirical methods in natural language processing, pp 327–335
33. Titov I, McDonald R (2008) Modeling online reviews with multi-grain topic models. In: Proceedings of the 17th international conference on World Wide Web, pp 111–120
34. Trabelsi A, Zaïane OR (2014) Mining contentious documents using an unsupervised topic model based approach. In: Proceedings of the IEEE international conference on data mining, pp 550–559
35. van Eemereen FH (2001) Crucial concepts in argumentation theory. Amsterdam University Press, Amsterdam
36. Wagstaff K, Cardie C, Rogers S, Schrödl S (2001) Constrained k-means clustering with background knowledge. In: Proceedings of the eighteenth international conference on machine learning, pp 577–584
37. Zhao WX, Jiang J, Yan H, Li X (2010) Jointly modeling aspects and opinions with a maxent-lda hybrid. In: Proceedings of the conference on empirical methods in natural language processing, pp 56–65

Amine Trabelsi is a Ph.D. student in Computing Science at the University of Alberta, Canada. He is a member of the Alberta Innovates Center for Machine Learning (AICML). He holds a M.Sc. in Computer Science from the University of Montreal, Canada, obtained in 2010. Amine received his B.Sc. in Computer Science and Management in 2007 from the University of Tunis, Tunisia. His research interests relate to Text Mining, Topic Modeling, Opinion Mining, Information Extraction, Data Mining and Machine Learning.

Osmar R. Zaïane is a Professor in Computing Science at the University of Alberta, Canada, and Scientific Director of the Alberta Innovates Centre for Machinre Learning (AICML). Dr. Zaiane joined the University of Alberta in July of 1999. He obtained a Master's degree in Electronics at the University of Paris, France, in 1989 and a Master's degree in Computer Science at Laval University, Canada, in 1992. He obtained his Ph.D. from Simon Fraser University, Canada, in 1999 under the supervision of Dr. Jiawei Han. His Ph.D. thesis work focused on web mining and multimedia data mining. He has research interests in data analytics, namely novel data mining algorithms, web mining, text mining, image mining, social network analysis, data visualization and information retrieval with applications in Health Informatics, e-Learning and e-Business. He has published more than 200 papers in refereed international conferences and journals, and taught on all six continents. Osmar Zaiane was from 2009 to 2012 the Secretary-Treasurer of the ACM SIGKDD (ACM Special Interest Group on Data Mining) which runs the world's premier data science, big data, and data mining association and conference. He is also on the steering committee of many data mining conferences such as IEEE International Conference on Data Mining, Advanced Data Mining and Applications, Data Science and Advanced Analytics. He was the Associate Editor then Editor-in-Chief of the ACM SIGKDD Explorations from 2003 to 2010. He is also Associate Editor of the Knowledge and Information Systems, An International Journal, by Springer, and of the journal Data Mining and Knowledge Discovery by Springer, as well as the International Journal of Internet Technology and Secured Transactions He was the General co-Chair of the IEEE International Conference on Data Mining ICDM 2011. Osmar Zaiane received the ICDM Outstanding Service Award in 2009 and the 2010 ACM SIGKDD Service Award.