Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019)

PhAITV: A Phrase Author Interaction Topic Viewpoint Model for the Summarization of Reasons Expressed by Polarized Stances

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

Abstract

This work tackles the problem of unsupervised modeling and extraction of the main contrastive sentential reasons conveyed by divergent viewpoints in text documents. It proposes a pipeline framework that is centered around the detection and clustering of phrases, assimilated to argument facets using a novel Phrase Author Interaction Topic-Viewpoint (PhAITV) model. The evaluation is conducted on all the components of the framework. It is mainly based on the informativeness, the relevance and the clustering accuracy of extracted reasons. The framework shows a significant improvement over several configurations and state-of-the-art methods in contrastive summarization on online debate datasets.

Introduction

Online debate forums provide a valuable resource for textual discussions about contentious issues. Contentious issues are controversial topics or divisive entities that usually engender opposing stances or viewpoints. Forum users write posts to defend their standpoint using persuasion, reasons or arguments. Such posts correspond to what we describe in (Trabelsi and Zaiane 2014; Trabelsi and Zaïane 2015) as contentious documents. An automatic tool that provides a contrasting overview of the main viewpoints and reasons given by opposed sides, debating an issue, can be useful for journalists and politicians. It provides them with systematic summaries and drafting elements on argumentation trends. In this work, given online forum posts about a contentious issue, we study the problems of unsupervised modeling and extraction, in the form of a digest table, of the main contrastive reasons conveyed by divergent viewpoints. Table 1 presents an example of a targeted solution in the case of the issue of "Abortion". The digest Table 1 is displayed à la ProCon.org or Debatepedia websites, where the viewpoints or stances engendered by the issue are separated into two columns. Each cell of a column contains an argument facet label followed by a sentential reason example. A sentential reason example is one of the infinite linguistic variations used to express a reason. For instance, the sentence "that cluster of cell is not a person" and the sentential reason "fetus is not a human" are different realizations of the same reason. For convenience, we will also refer to a sentence realizing a reason as a reason. **Reasons** in Table 1 are short sentential excerpts, from forum posts, which explicitly or implicitly express premises or arguments supporting a viewpoint. They correspond to any kind of intended persuasion, even if it does not contain clear argument structures (Habernal and Gurevych 2017). It should make a reader easily infer the viewpoint of the writer. **An argument facet** is an abstract concept corresponding to a low level issue or a subject that frequently occurs within arguments in support of a stance or in attacking and rebutting arguments of opposing stance (Misra et al. 2015). Similar to the concept of reason, many phrases can express the same facet. Phrases in bold in Table 1 correspond to **argument facet labels**, i.e., possible expressions describing argument facets. Reasons can also be defined as realizations of facets according to a particular viewpoint perspective. For instance, argument facet 4 in Table 1 frequently occurs within holders of Viewpoint 1 who oppose abortion. It is realized by its associated reason. The same facet is occurring in Viewpoint 2, in example 9, but it is expressed by a reason rebutting the proposition in example 4. Thus, reasons associated with divergent viewpoints can share a common argument facet. Exclusive facets emphasized by one viewpoint's side, much more than the other, may also exist (see example 5 or 8 in Table 1). Note that in many cases the facet label is very similar to the reason or proposition initially put forward by a particular viewpoint side, see examples 2 and 6, 7 in Table 1. It can also be a general aspect like "Birth Control" in example 5.

This paper describes the unsupervised extraction of these argument facets phrases and their exploitation to generate the associated sentential reasons in a contrastive digest table of the issue. Our first hypothesis is that detecting the main facets in each viewpoint leads to a good extraction of relevant sentences corresponding to reasons. Our second hypothesis is that leveraging the reply-interactions in online debate helps us cluster the viewpoints and adequately organize the reasons.

We distinguish three common characteristics of online debates, identified also by (Hasan and Ng 2014) and (Boltužić and Šnajder 2015), which make the detection and the clustering of argumentative sentences a challenging task. First, the unstructured and colloquial nature of used language makes it difficult to detect well-formed arguments. It makes it also

View 1 *Oppose*			*View 2* Support		
	Argument Facet Label	**Reason**		**Argument Facet Label**	**Reason**
1	**Fetus is not human**	What makes a fetus not human?	6	**Fetus is not human**	Fetus is not human
2	**Kill innocent baby**	Abortion is killing innocent baby	7	**Right to her body**	Women have a right to do what they want with their body
3	**Woman's right to control her body**	Does prostitution involves a woman's right to control her body?	8	**Girl gets raped and gets pregnant**	If a girl gets raped and becomes pregnant does she really want to carry that man's child?
4	**Give her child up for adoption**	Giving a child baby to an adoption agency is an option if a woman isn't able to be a good parent	9	**Giving up a child for adoption**	Giving the child for adoption can be just as emotionally damaging as having an abortion
5	**Birth control**	Abortion shouldn't be a form of birth control	10	**Abortion is not a murder**	Abortion is not a murder

Table 1: Contrastive Digest Table for Abortion.

noisy, containing non-argumentative portions and irrelevant dialogs. Second, the use of non-assertive speech acts like rhetorical questions to implicitly express a stance or to challenge opposing argumentation, like examples 1,3 and 8 in Table 1. Third, the similarity in words' usage between facet-related opposed arguments leads clustering to errors. Often a post rephrases the opposing side's premise while attacking it (see example 9). Note that exploiting sentiment analysis solely, like in product reviews, cannot help distinguishing viewpoints. Indeed, (Mohammad, Sobhani, and Kiritchenko 2017) show that both positive and negative lexicons are used, in contentious text, to express the same stance. Moreover, opinion is not necessarily expressed through polarity sentiment words, like example 6 in Table 1.

In this work, we do not explicitly tackle or specifically model the above-mentioned problems in contentious documents. However, we propose a generic data driven and facet-detection guided approach joined with posts' viewpoint clustering. It leads to extracting meaningful contrastive reasons and avoids running into these problems. Our contributions consist of: (1) the conception and deployment of a novel unsupervised generic pipeline framework producing a contrastive digest table of the main sentential reasons expressed in a contentious issue, given raw unlabeled posts from debate forums; (2) the devising of a novel Phrase Author Interaction Topic Viewpoint model, which jointly processes phrases of different length, instead of just unigrams, and leverages the interaction of authors in online debates; (3) the conduct of an extensive evaluation of the framework's final table output on real and noisy unstructured posts on different issues. The evaluation procedure of the proposed pipeline is conducted on the different components of the framework. It is mainly based on three measures of the final output: the informativeness of the digest as a summary, the relevance of extracted sentences as reasons and the accuracy of their viewpoint clustering. The results on different issues show that our model improves significantly over state-of-the-art methods and several baselines in terms of documents' summarization, reasons' retrieval and unsupervised contrastive reasons clustering.

Related Work

The objective of argument mining is to automatically detect the theoretically grounded argumentative structures within the discourse and their relationships (Stab and Gurevych 2014; Park and Cardie 2014). In this work, we are not interested in recovering the argumentative structures but, instead, we aim to discover the underpinning reasons behind people's opinion from online debates. In this section, we briefly describe some of the argument mining work dealing with social media text and present a number of important studies on Topic-Viewpoint Modeling.

The work on online discussions about controversial issues leverages the interactive nature of these discussions. Habernal and Gurevych (2017) consider rebuttal and refutation as possible components of an argument. Boltužić and Šnajder (2014) classify the relationship in a comment-argument pair as an attack (comment attacks the argument), a support or none. The best performing model of Hasan and Ng (2014)'s work on Reason Classification (RC) exploits the reply information associated with the posts. Most of the computational argumentation methods, including those mentioned above, are supervised. Moreover, the studies focusing on argument identification (Swanson, Ecker, and Walker 2015; Misra et al. 2017) , usually, rely on predefined lists of manually extracted arguments. As a first step towards unsupervised identification of prominent arguments from online debates, Boltužić and Šnajder (2015) group argumentative statements into clusters assimilated to arguments. However, only selected argumentative sentences are used as input. In this paper, we deal with raw posts containing both argumentative and non-argumentative sentences.

Topic-Viewpoint models are extensions of Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) applied to contentious documents. They hypothesize the existence of underlying topic and viewpoint variables that influence the author's word choice when writing about a controversial issue. The viewpoint variable is also called stance, perspective or argument variable in different studies. Topic-Viewpoint models are mainly data-driven approaches which reduce the documents into topic-viewpoint dimensions. A

Topic-Viewpoint pair t-v is a probability distribution over unigram words. The unigrams with top probabilities characterize the used vocabulary when talking about a specific topic t while expressing a particular viewpoint v at the same time. Several Topic-Viewpoint models of controversial issues exist (Qiu and Jiang 2013; Trabelsi and Zaiane 2014; Thonet et al. 2016). Little work is done to exploit these models in order to generate sentential digests or summaries of controversial issues instead of just producing distributions over unigram words. Below we introduce the research that is done in this direction.

Paul, Zhai, and Girju (2010) are the first to introduce the problem of contrastive extractive summarization on online surveys and editorials data. They propose the Topic Aspect Model (TAM) and use its output distributions to compute similarity scores between sentences. Comparative LexRank, a modified LexRank (Erkan and Radev 2004), is run on scored sentences to generate the summary. Recently,Vilares and He (2017) propose a topic-argument or viewpoint model called the Latent Argument Model_LEX (LAM_LEX). Using LAM_LEX, they generate a succinct summary of the main viewpoints from a parliamentary debates dataset. The generation consists of ranking the sentences according to a discriminative score for each topic and argument dimension. It encourages higher ranking of sentences with words exclusively occurring with a particular topic-argument dimension which may not be accurate in extracting the contrastive reasons sharing common words. Both of the studies, cited above, exploit the unigrams output of their topic-viewpoint modeling.

In this work, we propose a Topic-Viewpoint modeling of phrases of different length, instead of just unigrams. We believe phrases allow a better representation of the concept of argument facet. They would also lead to extract a more relevant sentence realization of this latter. Moreover, we leverage the interactions of users in online debates for a better contrastive detection of the viewpoints.

The Pipeline Methodology

Phrase Mining Phase

The inputs of this module are raw posts (documents). We prepare the data by removing identical portions of text in replying posts. We also delete entirely duplicated posts. We remove stop and rare words. We consider working with the stemmed version of the words. The objective of the phrase mining module is to partition the documents into high quality bag-of-phrases instead of bag-of-words. Phrases are of different length, single or multi-words. We follow the steps of El-Kishky et al. (2014), who propose a phrase extraction procedure for the Phrase-LDA model. Given the contiguous words of each sentence in a document, the phrase mining algorithm employs a bottom-up agglomerative merging approach. At each iteration, it merges the best pair of collocated candidate phrases if their statistical significance score exceeds a threshold which is set empirically (according to El-Kishky et al. (2014)'s implementation). The significance score depends on the collocation frequency of candidate phrases in the corpus. It measures their number of standard deviation away from the expected occurrence under an independence null hypothesis. The higher the score, the more likely the phrases co-occur more often than by chance.

Topic-Viewpoint Modeling Phase

In this section, we present the Phrase Author Interaction Topic-Viewpoint model (PhAITV). It takes as input the documents, partitioned in high quality phrases of different lengths, and the information about author-reply interactions in an online debate forum. The objective is to assign a topic and a viewpoint label to each occurrence of the phrases. This would help to cluster them into Topic-Viewpoint classes. We assume that A authors participate in a forum debate about a particular issue. Each author a writes D_a posts. Each post d_a is partitioned into G_{da} phrases of different lengths (>=1). Each phrase contains M_{gda} words. Each term w_{mg} in a document belongs to the corpus vocabulary of distinct terms of size W. In addition, we assume that we have the information about whether a post replies to a previous post or not. Let K be the total number of topics and L be the total number of viewpoints. Let θ_{da} denote the probability distribution of K topics under a post d_a; ψ_a be the probability distributions of L viewpoints for an author a; ϕ_{kl} be the multinomial probability distribution over words associated with a topic k and a viewpoint l; and ϕ_B a multinomial distribution of background words. The generative process of a post according to the PhAITV model (see Figure 1) is the following. An author a chooses a viewpoint v_{da} from the distribution ψ_a. For each phrase g_{da} in the post, the author samples a binary route variable x_{gda} from a Bernoulli distribution σ. It indicates whether the phrase is a topical or a background word. Multi-word phrases cannot belong to the background class. If $x_{gda} = 0$, the word is sampled from ϕ_B. Otherwise, the author, first, draws a topic z_{gda} from θ_{da}, then, samples each word w_{mg} in the phrase from the same $\phi_{z_{gda}v_{da}}$.

Note that, in what follows, we refer to a current post with index id and to a current phrase with index ig. When the current post is a reply to a previous post by a different author, it may contain a rebuttal or it may not. If the reply attacks the previous author then the rebuttal variable Rb_{id} is set to 1 else if it supports, the rebuttal takes 0. We define the **parent posts** of a current post as all the posts of the author who the current post is replying to. Similarly, the **child posts** of a current post are all the posts replying to the author of the current post. We assume that the probability of a rebuttal $Rb_{id} = 1$ depends on the degree of opposition between the viewpoint v_{id} of the current post and the viewpoints $\mathcal{V}_{id}^{par}$ of its parent posts as the following:

$$p(Rb_{id} = 1 | v_{id}, \mathcal{V}_{id}^{par}) = \frac{\sum_{l'}^{\mathcal{V}_{id}^{par}} \mathbf{I}(v_{id} \neq l') + \eta}{|\mathcal{V}_{id}^{par}| + 2\eta}, \quad (1)$$

where $\mathbf{I}(condition)$ equals 1 if the condition is true and η a smoothing parameter.

For the inference of the model's parameters, we use the collapsed Gibbs sampling. For all our parameters, we set fixed symmetric Dirichlet priors. According to Figure 1, the

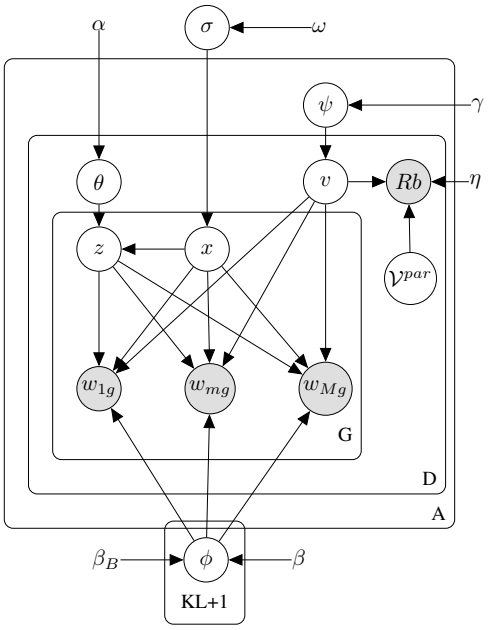

Figure 1: Plate Notation of The PhAITV model

Rb variable is observed. However, the true value of the rebuttal variable is unknown to us. We fix it to 1 to keep the framework purely unsupervised, instead of estimating the reply disagreement using methods based on lexicon polarity. Setting $Rb = 1$ means that all replies of any post are rebuttals attacking all of the parent posts excluding the case when the author replies to his own post. This comes from the observation that the majority of the replies, in the debate forums framework, are intended to attack the previous proposition (Hasan and Ng 2013). This setting will affect the viewpoint sampling of the current post. The intuition is that, if an author is replying to a previous post, the algorithm is encouraged to sample a viewpoint which opposes the majority viewpoint of parent posts (Equation 1). Similarly, if the current post has some child posts, the algorithm is encouraged to sample a viewpoint opposing the children's prevalent stance. If both parent and child posts exist, the algorithm is encouraged to oppose both, creating some sort of adversarial environment when the prevalent viewpoints of parents and children are opposed. The derived sample equation of current post's viewpoint v_{id} given all the previous sampled assignments in the model $\vec{v}_{\neg id}$ is:

$$p(v_{id} = l|\vec{v}_{\neg id}, \vec{w}, \vec{Rb}, \vec{x}) \propto$$
$$n^{(l)}_{a,\neg id} + \gamma \times \frac{\prod_{t}^{W_{id}} \prod_{j=0}^{n^{(t)}_{id}-1} n^{(t)}_{l,\neg id} + j + \beta}{\prod_{j=0}^{n_{id}-1} n^{(.)}_{l,\neg id} + W\beta + j} \times p(Rb_{id} = 1|v_{id}, \mathcal{V}^{par}_{id}) \times \prod_{c|v_{id}\in\mathcal{V}^{par}_{c}} p(Rb_c = 1|v_c, \mathcal{V}^{par}_{c}). \quad (2)$$

The count $n^{(l)}_{a,\neg id}$ is the number of times viewpoint l is assigned to author a's posts excluding the assignment of current post, indicated by $\neg id$; $n^{(t)}_{l,\neg id}$ is the number of times term t is assigned to viewpoint l in the corpus excluding assignments in current post; $n^{(.)}_{l,\neg id}$ is the total number of words assigned to l; W_{id} is the set of vocabulary of words in post id; $n^{(t)}_{id}$ is the number of time word t occurs in the post. The third term of the multiplication in Equation 2 corresponds to Equation 1 and is applicable when the current post is a reply. The fourth term of the multiplication takes effect when the current post has child posts. It is a product over each child c according to Equation 1. It computes how much would the children's rebuttal be probable if the value of v_{id} is l. Given the assignment of a viewpoint $v_{id} = l$, we also jointly sample the topic and background values for each phrase ig in post id, according to the following:

$$p(z_{ig} = k, x_{ig} = 1|\vec{z}_{\neg ig}, \vec{x}_{\neg ig}, \vec{w}, \vec{v}) \propto \prod_{j=0}^{M_{ig}} n^{(1)}_{\neg ig} + \omega + j \times n^{(k)}_{id,\neg ig} + \alpha + j \times \frac{n^{(w_{jg})}_{kl,\neg ig} + \beta}{n^{(.)}_{kl,\neg ig} + W\beta + j}, \quad (3)$$

$$p(x_{ig} = 0|\vec{x}_{\neg ig}, \vec{w}) \propto \prod_{j=0}^{M_{ig}} n^{(0)}_{\neg ig} + \omega + j \times \frac{n^{(w_{jg})}_{0,\neg ig} + \beta_B}{n^{(.)}_{0,\neg ig} + W\beta_B + j}. \quad (4)$$

Here $n^{(k)}_{id,\neg ig}$ is the number of words assigned to topic k in post id, excluding the words in current phrase ig; $n^{(1)}_{\neg ig}$ and $n^{(0)}_{\neg ig}$ correspond to the number of topical and background words in the corpus, respectively; $n^{(w_{jg})}_{kl,\neg ig}$ and $n^{(w_{jg})}_{0,\neg ig}$ correspond to the number of times the word of index j in the phrase g is assigned to topic-viewpoint kl or is assigned as background; $n^{(.)}$s are summations of last mentioned expressions over all words.

After the convergence of the Gibbs algorithm, each multi-word phrase is assigned a topic k and a viewpoint l label. We exploit these labels to first create clusters, where each cluster corresponds to a topic-viewpoint value kl. It contains all the phrases that are assigned to kl at least one time. Second, we rank the phrases inside each cluster according to their assignment frequencies.

Grouping and Facet Labeling Phase

The inputs of this module are Topic-Viewpoint clusters, each containing ranked multi-word phrases, along with their frequency scores. The outputs are sorted phrases corresponding to argument facet labels for each viewpoint. This phase is based on two assumptions. (1) Grouping constructs agglomerations of lexically related phrases. which can be assimilated to the notion of argument facets. (2) An argument facet is better expressed with a Verb Expression than a Noun Phrase. A Verbal Expression (VE) is a sequence of correlated chunks centered around a Verb Phrase chunk (Li et al.

2015). We believe that encouraging labeling an argument facet with a VE, over a Noun Phrase, reduces the search space for the sentential reasons and makes the extraction more accurate.

We propose a second layer of phrase grouping on each of the input Topic-Viewpoint clusters. It is based on the number of words overlap between stemmed pairs of phrases. The number of groups is not a parameter. First, we compute the number of words overlap between all pairs and sort them in descending order. Then, while iterating on them, we encourage a pair with matches to create its own group if both of its phrases are not grouped yet. If it has only one element grouped, the other element joins it. If a pair has no matches, then each non-clustered phrase creates its own group.

Some of the generated groups may contain small phrases that can be fully contained in longer phrases of the same group. We remove them and add up their scores to corresponding phrases. If there is a conflict where two or several phrases can contain the same phrase, then the one that is a Verbal Expression adds up the number of assignments. This procedure inflates the assignment score of VE phrases in order to promote them to be solid candidates for the argument facet labeling. The final step consists of collecting the groups pertaining to each Viewpoint, regardless of the topic, and sort them based on the cumulative score of their composing phrases. This will create viewpoint clusters with groups which are assimilated to argument facets. The labeling consists of choosing one of the phrases as the representative of the group. We simply choose the one with the highest frequency score to obtain Viewpoint clusters of argument facet labels.

Extraction of Contrastive Reasons Phase

The inputs of this final phase are sorted facet phrases for each Viewpoint, plus, all the sentences containing these phrases' original viewpoint and topic assignments according to PhAITV. The output is the digest table of contrastive reasons. In order to extract a short sentential reason, given the phrase label and corresponding sentences, we follow these steps: (1) find the set of sentences with the most common overlapping words among all the sentences, disregarding the set of words composing the facet label; (2) choose the shortest sentence in the set. The process is repeated for all sorted phrases, according to the desired number of sentences to display in the digest table. Note that duplicate sentences within the Viewpoint are removed. Also, we restore stop and rare words of the phrases when rendering them as argument facets similar to those in Table 1. We choose the most frequent sequence in related sentences.

Experiments and Results

We, first, present the used datasets then, we evaluate the different components of our proposed framework and the final extracted sentential reasons according to their informativeness, their relevance and the accuracy of their viewpoint clustering. We perform a qualitative evaluation of the generated argument facets. However, a direct quantitative evaluation of the argument facets and their labels is not the objective of this work. The final digest is dependent on the facets' generation. Thus, facets are evaluated indirectly by assessing the subsequent sentential reasons digest.

Datasets

We exploit the reasons corpus constructed by Hasan and Ng (2014). We consider the issues of Abortion and Gay Rights. The posts are extracted from CreateDebate forum. Each post has a stance label (i.e., support or oppose the issue). The argumentative sentences of the posts are labeled with a reason label from a set of predefined reason labels associated with each stance. The reason labels can be assimilated to argument facets. The dataset contains 13 labels for Abortion and 9 for Gay Rights. The number of posts are 1876 for Abortion and 1363 for Gay Rights. Only a subset of the posts, for each dataset, has its sentences annotated with reasons according to (Hasan and Ng 2014). The argumentative sentence percentage in this subset is 20.4 and 29.8, for Abortion and Gay Rights, respectively. The percentage of disagreeing or rebuttal replies is 67.05 for Abortion and 66.61 for Gay Rights. The PhAITV model exploits only the text, the author identities and the information about whether a post is a reply or not. For evaluation purposes, we leverage the subset of argumentative sentences which is annotated with reasons labels to construct several reference summaries for each dataset. Each reference summary contains a combination of sentences, each from one possible label (13 for Abortion, 9 for Gay Rights). This makes the references exhaustive and reliable resources on which we can build a good measure of informativeness. The number of reference summaries is 100.

Experiments Set Up

Throughout the experiments we evaluate both the intermediary and final outputs of the proposed pipeline framework. Our framework is composed of a Phrase Mining phase, a Topic-Viewpoint modeling (PhAITV), a Grouping and labeling and a final Table Extraction phase. We refer to this combination as "***PhAITV + Grouping + Extraction***". In the following sections, we assess the final summary table produced by this setting with different settings of the framework, along with similar state-of-the-art methods. The objective is to demonstrate the importance of the different components and show that the proposed "*PhAITV + Grouping + Extraction*" outperforms existing contrastive summarization approaches.

In order to evaluate the Phrase Mining phase, we propose a degenerated unigram version of PhAITV, **AITV**. AITV is described in details in our previous work (Trabelsi and Zaïane 2018). In AITV based setting, no grouping is involved and the query of retrieval consists of the top three keywords instead of the phrase. In order to evaluate Topic Viewpoint Modeling, we propose to substitute PhAITV with **PhJTV**, an augmented phrase version of one of our previous Topic-Viewpoint model, **JTV** (Trabelsi and Zaïane 2016). JTV is a unigram Topic-Viewpoint model that has demonstrated effectiveness in generating Topic-Viewpoint word dimension comparing to LDA when using constrained clustering. We also explore a modified setting of the framework, "***PhAITV + Extraction***", where the grouping component is ignored. Similarly, we try "***PhAITV + Grouping +***

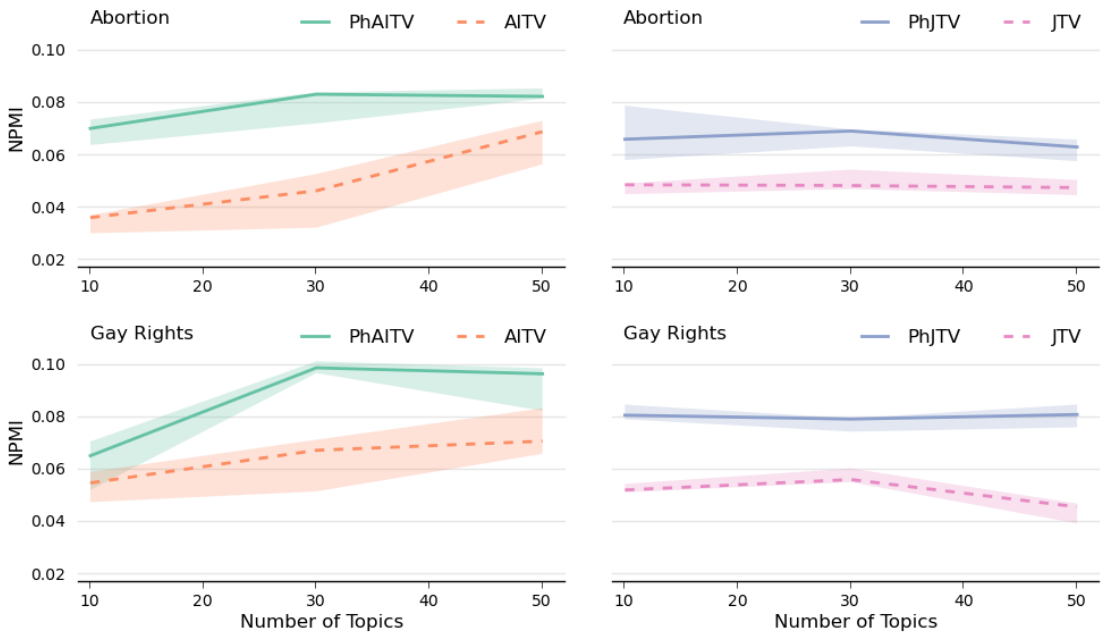

Figure 2: Medians and quartiles of average NPMI on the outputs of PhAITV, AITV, PhJTV and JTV for Abortion and GayRights

LexRank", where we replace the Extraction procedure with the LexRank algorithm (Erkan and Radev 2004), to rank sentences and choose the one with the highest score as a sentential reason. We also compare against two state-of-the-art studies in generating contrastive summarization from contentious text in general, which are generic enough to not depend on the structure of the data. These correspond to Paul, Zhai, and Girju (2010)'s work and recent Vilares and He (2017)'s study. They are based on Topic-Viewpoint models, **TAM** (Paul, Zhai, and Girju 2010), and **LAM_LEX** (Vilares and He 2017) (see Section Related Work). Below, we refer to the names of these two Topic-Viewpoint methods to describe the whole process that produces their final summary. As a weak baseline, we generate **random summaries** from the set of possible sentences in each corpus. We also create **correct summaries** from the subset of labeled argumentative sentences. Moreover, we compare with another version of our framework, including **PhAITV$_{\text{view}}$**, which assumes the true values of the posts' viewpoints are given.

In the Phrase mining phase, the parameters are set similar to El-Kishky et al. (2014). We try different combinations of the PhAITV's hyperparameters and use the combination which gives a satisfying overall performance. During experiments, we did not observe a significant change in performance when the hyperparameters were varied. PhAITV's hyperparameters are set as follows: $\alpha = 0.1$; $\beta = 1$; $\gamma = 1$; $\beta_B = 0.1$; $\eta = 0.01$; $\omega = 10$. The number of the Gibbs Sampling iterations is 1500. The number of viewpoints L equals 2. We try a different number of topics K for each Topic-Viewpoint model used in the evaluation. For each model, the chosen K achieves a satisfying NPMI coherence score on the two datasets (See Figure 2 as examples). The values of K are set to 30,50,30,30,10 and 10 for PhAITV, AITV, PhJTV, JTV LAM_LEX, and TAM, respectively. Other parameters of the methods used in the comparison are set to their default values. All the models generate their top 15 sentences for Abortion and their 10 best sentences for Gay Rights. These are rounded values of the numbers of reason labels (corresponding to the number of sentences in reference summaries) (see Section Datasets).

Evaluation of the Phrase Topic Viewpoint Modeling

In this section, we evaluate the intermediary output of the combined Phrase Mining and Topic Viewpoint modeling phases of the framework. In particular, we assess the coherence of the 10 distinct words of the top phrases of each cluster $\mathcal{P}_{kl}$ produced by PhAITV for each Topic-Viewpoint kl. We compare the coherence of PhAITV output to that of the degenerated version AITV, and do the same for PhJTV and JTV. In order to automatically measure the coherence of Topic Viewpoint models, we use the average Normalized Pointwise Mutual Information (NPMI) (Bouma 2009) between pairs of the top 10 words in each Topic-Viewpoint cluster. This measure correlates well with human evaluations on topics' coherence (Aletras and Stevenson 2013; Lau, Newman, and Baldwin 2014). An NPMI between two words is function of their co-occurrence probabilities in the corpus. It takes a maximum of 1 when the words only occur together, and a minimum of -1 when they never co-occur.

Figure 2 presents median and quartile values of average NPMI, measured on the outputs of PhAITV, AITV, PhJTV and JTV, and aggregated over 5 runs for different number of topics {10,30,50}, using Abortion and GayRights datasets. We observe that the models with a phrase mining module, PhAITV and PhJTV, significantly outperform their

Figure 3: Word Clouds of argument facet labels generated by "*PhAITV + Grouping + Extraction*" (left) and "*PhAITV + Extraction*" (right)

corresponding unigram models, AITV and JTV, in terms of top-words coherence, for both datasets. This confirms the assumption that using the phrase mining module yields more coherent Topic-Viewpoint dimensions than considering only unigrams. However, this does not necessarily mean that phrase models lead to a better extraction of sentential reasons. We examine the effect of phrase mining on the extraction of relevant and informative sentential reason in the following sections. In general, PhAITV reaches higher median NPMI values than PhJTV. We will later compare the final output of the pipeline framework in terms of reasons clustering when using each one of these models as a Topic-Viewpoint component. Note that separately evaluating the performance of our Topic Viewpoint model in terms of document clustering has shown satisfiable results which we do not report here for lack of space. We are more interested in its impact on the final sentence-level viewpoint clustering.

Qualitative Evaluation of Grouping and Labeling on Abortion

In this section, we qualitatively evaluate the outputs of the Grouping and Labeling phase. The objective is to qualitatively verify the assumption that the output phrases of this module are effectively labeling argument facets. An argument facet is described, in the introduction section, as an abstract concept corresponding to a sub-issue or a theme that frequently occurs within arguments in support of a viewpoint or in attacking arguments of opposing stance. As an abstract notion, an argument facet can be expressed or labeled with different expressions. Figure 3 presents a word cloud of the phrase labels generated by our framework "*PhAITV + Grouping + Extraction*" exploiting Grouping and Labeling (left side cloud), and another cloud produced by "*PhAITV + Extraction*", the version without the Grouping and Labeling (right side cloud). For each variant, the cloud is generated from the top phrases of three digest-tables on Abortion. Bigger font phrases are reoccurring more often across the tables. We observe that Grouping and Labeling module generates precise and self-contained phrases that correspond to the common argument facets expressed in the issue of Abortion (see Hasan and Ng (2014)'s reasons labels on Abortion). The phrases produced by the non-Grouping version can also represent argument facets, however they are not as precise as those of Grouping version. They seem more general (e.g., taking human life Vs. human life). Precision is needed to narrow the search space for relevant sentences in the extraction module. Most of the left-side phrases are Verbal Expressions while most of the right side ones are Noun phrases. Thus, encouraging verbal expressions in Grouping and Labeling phase plays a role in obtaining good labels of argument facets. The diversity and recall inside the left side cloud is higher than on the right side (e.g., mother's life in danger, putting the child up for adoption). This is the consequence of grouping the lexically similar phrases. The grouping allows to avoid repetitiveness, and, thus, is more likely to generate diverse phrases. This diversity of argument facets will reflect on the extracted sentential reasons. This can be observed in the sample sentential reason's output in Table 4.

Evaluation of Digest Table Informativeness

The remaining sections evaluate the quality of the final sentential reasons digest table according to different criteria. An example of sentential reasons digest table produced by our framework "PhAITV + Grouping + Extraction" is displayed in Table 4 . On each criterion, we compare "PhAITV + Grouping + Extraction" against several variants in order to assess the contribution of each module. Moreover, an evaluation of our proposed pipeline should take into account all of the three criteria, i.e., informativeness, relevance and viewpoint clustering of the sentential reasons, because of their complementarity. Here we re-frame the problem of creating a contrastive digest table into a summary problem. The concatenation of all extracted sentential reasons of the digest is considered as a candidate summary. The construction of reference summaries is explained in Datasets Section. It favors the diversity within the references. Informativeness denotes the degree to which a candidate summary is similar to exhaustive reference summaries. The more similar to

	Gay Rights			Abortion		
	R2-R	R2-P	R2 F-M	R2-R	R2-P	R2 F-M
Random Summaries	0.7	0.9	0.8	1.0	1.0	1.0
Correct Summaries	3.0	3.0	3.0	5.8	5.1	5.4
JTV (Trabelsi and Zaïane 2016) + Extraction	2.5	2.1	2.3	3.4	2.8	3.1
PhJTV + Grouping + Extraction	2.5	3.0	2.7	4.2	4.3	4.2
AITV (Trabelsi and Zaïane 2018) + Extraction	2.7	3.0	2.8	3.0	2.6	2.8
PhAITV + Grouping + Extraction	2.7	3.0	2.8	4.5	**4.7**	**4.6**
PhAITV + Extraction	**2.9**	**3.2**	**3.0**	3.1	3.6	3.3
PhAITV + Grouping + LexRank	2.8	2.8	2.8	**5.0**	3.7	4.2
TAM (Paul, Zhai, and Girju 2010)	2.0	3.1	2.4	1.8	2.4	2.1
LAM_LEX (Vilares and He 2017)	1.1	0.8	0.9	1.5	0.8	1.0

Table 2: Averages of ROUGE-2 Measures (in %, stemming and stop words removal applied) on Gay Rights and Abortion (values in "bold" represent best values disregarding Correct Summaries values)

the reference, the more exhaustive the candidate summary. We evaluate all competing methods using ROUGE evaluation metric (Lin 2004), a measure often used for automatic summaries evaluation. We report the results of Rouge-2's Recall (R-2 R), Precision (R-2 P) and F-Measure (R-2 F-M). Rouge-2 captures the similarities between sequences of bigrams in references and candidates. The higher the measure, the more similar to the reference, the summary is. All reported ROUGE-2 values are computed after applying stemming and stop words removal on reference and candidate summaries. This procedure may also explain the relatively small values of reported ROUGE-2 measures in Table 2, compared to those usually computed when stop words are not removed. The existence of stop words in candidate and references sentences increases the overlap, and hence the ROUGE measures' values in general. Applying stemming and stop words removal was based on some preliminary tests that we conducted on our dataset. The tests showed that two candidate summaries containing different numbers of valid reasons, would have a statistically significant difference in their ROUGE-2 values when stemming and stop words removal are applied.

Table 2 contains the averaged results, over 10 generated summaries, on Abortion and Gay Rights, respectively. We observe that all degenerate versions of our framework produce significantly better summaries than the weak Random Summaries baseline. Their ROUGE values are comparable to those of the correct summaries on Gay Rights. All PhAITV-based versions produce more informative summaries than their unigram-based counterpart AITV, on Abortion. Summaries are comparable on Gay Rights. The same pattern is observed with JTV based configuration of our framework and its enhanced PhJTV version. This confirms the assumption that exploiting phrases rather than unigram models within our framework can lead to more informative summaries. The difference between the summaries of PhAITV-based and PhJTV-based settings, or between AITV and JTV, in terms of ROUGE-2 metric is not significant. The difference between these models is better discerned on their ability to distinguish viewpoints (see Evaluation of Digest Table Relevance and Contrast Section and Table 3).

The PhAITV versions including grouping phase yield significantly better results, on Abortion, than the version without grouping. The non-grouping variant, however, has a slightly, but not significantly, better informative summaries on Gay Rights. The "*PhAITV + Grouping + LexRank*" variant has a better ROUGE-2 recall, on Abortion, than the proposed "*PhAITV + Grouping + Extraction*" . We believe this is due to the longer extracted sentences by LexRank compared to the conciseness restriction encoded in the extraction phase. Nonetheless, "*PhAITV + Grouping + Extraction*" gives better precision and F-Measure trade-offs.

The recent contrastive summarization approach LAM_LEX (Vilares and He 2017) performs poorly in this task (close to Random summaries) for both datasets. "*PhAITV + Grouping + Extraction*" performs significantly better than TAM on Abortion, and slightly better on Gay Rights. The output digests in Table 4 showcase the superiority of PhAITV framework compared to TAM and LAM_LEX. We notice that PhAITV's digest produces different types of reasons from diverse argument facets, like putting child up for adoption, life begins at conception, the religion argument, mother's life in danger. However, such informativeness on these different argument facets is lacking on both digests of LAM_LEX and TAM. For instance, we remark the recurrence of the subject of killing or taking human life with different sentences in TAM's digest. In terms of ROUGE measure, interestingly, the summaries of AITV configuration are more informative than similar unigram-based summaries of TAM and LAM_LEX, on both datasets. This suggests that the proposed pipeline is effective in terms of summarization even without the phrase modeling.

Evaluation of Digest Table Relevance and Contrast

For the following evaluations, we conducted a human annotation task with three annotators, after ethical approval. The annotators were acquainted with both studied issues and the possible reasons conveyed by each side. They were given lists of mixed sentences generated by the models. They were asked to indicate the stance of each sentence ((+) for, (-) against) when it contains any kind of persuasion, reason-

	Gay Rights			Abortion		
	Rel	NPV	Acc.	Rel	NPV	Acc.
JTV (Trabelsi and Zaïane 2016) + Extraction	0.60	45.00	44.44	0.66	47.62	45.45
PhJTV + Grouping + Extraction	**0.80**	50.00	46.42	0.90	50.00	46.15
AITV (Trabelsi and Zaïane 2018) + Extraction	0.50	**75.00**	66.66	0.66	58.33	59.09
PhAITV + Grouping + Extraction	**0.80**	**75.00**	**75.00**	**0.93**	**75.00**	**73.62**
PhAITV + Extraction	0.66	66.66	66.66	0.73	50.00	49.09
PhAITV + Grouping + LexRank	0.70	33.33	52.38	0.80	56.66	56.36
TAM (Paul, Zhai, and Girju 2010)	0.50	50.00	42.85	0.53	50.00	46.42
LAM_LEX (Vilares and He 2017)	0.50	50.00	50.00	0.40	50.00	64.44
$\text{PhAITV}_{\text{view}}$+ Grouping + Extraction	0.90	100.0	100.0	0.93	87.5	83.33

Table 3: Median values of Relevance Rate (Rel), Negative Predictive Value (NPV) and Clustering Accuracy Percentages (Acc.) on GayRights and Abortion (values in "bold" represent best values disregarding $\text{PhAITV}_{\text{view}}$ values)

ing or argumentation from which they could easily infer the stance. Thus, if they label the sentence, the sentence is considered a relevant reason. Otherwise, the sentence is not a reason and irrelevant (represented by (0)). The average Kappa agreement between the annotators was 0.66. The final annotations correspond to the majority label. In case of a conflict between the annotators, we consider the sentence irrelevant. We measure the Relevance (Rel.) by the ratio of the number of relevant sentences (judged as (+) or (-)) divided by the number of the digest's sentences.

Table 3 contains the median relevance (Rel) rates over 5 summaries, on GayRights and Abortion, respectively. Two main observations can be made : (1) all the phrase-based variants generate more relevant outputs than all of the unigram-based approaches, consolidating the idea that phrases leads to a better sentential reason retrieval; (2) the configurations achieving the best relevance rates are those following our proposed pipeline framework phrase+Grouping+Extraction. Furthermore, "*PhAITV + Grouping + Extraction*" realizes high relevance rates, comparable to those of the heavily guided $\text{PhAITV}_{\text{view}}$, and outperforming its rivals, TAM and LAM_LEX, by a very large margin on both datasets. This is also showcased by Table 4's examples. The ratio of sentences judged as reasons given to support a stance ((+) or (-)) is higher for PhAITV-based digest. Interestingly, even the PhAITV's sentences judged as irrelevant are not off-topic. They include relevant expressions like "abortion is murder" or "women might choose to abort", which are the corresponding argument facets leveraged for their extraction. They are also coherent with other sentences in the clusters in terms of viewpoint. It is important to note that sentences and argument facets presented earlier in Table 1 are also collected from our PhAITV + Grouping + Extraction outputs. Reasons 1, 3 and 8 reveal the ability of the system to display rhetorical questions.

All compared models generate sentences for each viewpoint. Given the human annotations, we consider assessing the viewpoint clustering of the relevant extracted sentences by two measures: the Clustering Accuracy and the Negative Predictive Value (NPV). NPV consider a pair of sentences as unit. It corresponds to the number of true stance opposed pairs in different clusters divided by the number of pairs formed by sentences in opposed clusters.

A high NPV is an indicator of a good inter-clusters opposition i.e., a good contrast of sentences' viewpoints. Table 3 reports the median NPV and Accuracy values over 5 generated summaries for each variant. A good viewpoint clustering of the sentential reasons depends on a good viewpoint assignment of the phrases and the documents. Thus, the performance depends on how well the Topic-Viewpoint modeling distinguishes the viewpoints. Table 3 shows that most of the PhAITV degenerate versions, including AITV, achieve better NPV and accuracy than JTV variants, TAM and LAM_LEX, on both datasets. This confirms the hypothesis that leveraging the reply-interactions, in online debate, helps detect the viewpoints of posts and subsequently correctly cluster the reasons' viewpoints. The proposed configuration "*PhAITV + Grouping + Extraction*" achieves very encouraging NPV and accuracy results without any supervision. Again, it outperforms significantly the state-of-the-art methods in unsupervised contrastive summarization. Table 4 shows a much better alignment, between the viewpoint clusters and the stance signs of reasons (+) or (-), for PhAITV comparing to competitors. The NPV and accuracy values of the sample digests are close to the median values reported in Table 3. The contrast also manifests when similar facets are discussed but by opposing viewpoints like in "life begins at conception" against "fetus before it can survive outside the mother's womb is not a person". The results are not close yet to $\text{PhAITV}_{\text{view}}$-based variant which achieves a 100% accuracy on Gay Rights and a 0.9 relevance rate.

Conclusion

This work proposes an unsupervised framework for the detection, clustering, and displaying of the main sentential reasons conveyed by divergent viewpoints in contentious text from online debate forums. A pipeline approach is suggested based on a Phrase Mining module and a novel Phrase Author Interaction Topic-Viewpoint model. The evaluation of the approach is based on three measures computed on the final digest: the informativeness, the relevance and the accuracy of viewpoint clustering. The results on contentious is-

PhAITV + Grouping + Extraction			
	Viewpoint 1		Viewpoint 2
(-)	If a mother or a couple does not want a child there is always the option of putting the child up for adoption.	(+)	The fetus before it can survive outside of the mother's womb is not a person.
(-)	I believe life begins at conception and I have based this on biological and scientific knowledge.	(+)	Giving up a child for adoption can be just as emotionally damaging as having an abortion.
(-)	God is the creator of life and when you kill unborn babies you are destroying his creations.	(+)	you will have to also admit that by definition; abortion is not murder.
(-)	I only support abortion if the mothers life is in danger and if the fetus is young.	(-)	No abortion is wrong.
(0)	The issue is whether or not abortion is murder.	(0)	I simply gave reasons why a woman might choose to abort and supported that.
LAM_LEX (Vilares and He 2017)			
	Viewpoint 1		Viewpoint 2
(-)	abortion is NOT the only way to escape raising a child that would remind that person of something horrible	(+)	if a baby is raised by people not ready, or incapable of raising a baby, then that would ruin two lives.
(+)	I wouldn't want the burden of raising a child I can't raise	(+)	The fetus really is the mother's property naturally
(0)	a biological process is just another name for metabolism	(0)	Now this is fine as long as one is prepared for that stupid, implausible, far-fetched, unlikely, ludicrous scenario
(0)	The passage of scripture were Jesus deals with judging doesn't condemn judging nor forbid it	(0)	you are clearly showing that your level of knowledge in this area is based on merely your opinions and not facts.
(0)	your testes have cells which are animals	(0)	we must always remember how life is rarely divided into discreet units that are easily divided
TAM (Paul, Zhai, and Girju 2010)			
	Viewpoint 1		Viewpoint 2
(-)	I think that is wrong in the whole to take a life.	(+)	Or is the woman's period also murder because it also is killing the potential for a new human being?
(-)	I think so it prevents a child from having a life.	(-)	it maybe then could be considered illegal since you are killing a baby, not a fetus, so say the fetus develops into an actuall baby
(+)	Abortion is not murder because it is performed before a fetus has developed into a human person.	(0)	In your scheme it would appear to be that there really is no such thing as the good or the wrong.
(0)	He will not obey us.	(0)	NO ONE! but God.
(0)	What does it have to do with the fact that it should be banned or not?	(0)	What right do you have to presume you know how someone will life and what quality of life the person might have?

Table 4: Sample Digest Tables Output of sentential reasons produced by the frameworks based on PhAITV, LAM_LEX and TAM when using Abortion dataset from CreateDebate. Sentences are labeled according to their stances as the following: (+) reason for abortion; (-) reason against abortion; and (0) irrelevant

sues show that PhAITV-based pipeline outperforms several baselines and state-of-the-art methods for each of these criteria. In this research, we dealt with contentious documents in online debate forums, which usually enclose a high rate of rebuttal replies. Other social media platforms, like Twitter, may not have rebuttals as common as in online debates. Moreover, a manual inspection of digest tables suggests the need for improvement in the detection of semantically similar reasons and their hierarchical clustering.

References

Aletras, N., and Stevenson, M. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics*, 13–22.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Boltužić, F., and Šnajder, J. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, 49–58. Baltimore, Maryland: Association for Computational Linguistics.

Boltužić, F., and Šnajder, J. 2015. Identifying prominent arguments in online debates using semantic textual similarity.

In *Proceedings of the 2nd Workshop on Argumentation Mining*, 110–115. Denver, CO: Association for Computational Linguistics.

Bouma, G. 2009. Normalized pointwise mutual information in collocation extraction. *Proceedings of GSCL* 31–40.

El-Kishky, A.; Song, Y.; Wang, C.; Voss, C. R.; and Han, J. 2014. Scalable topical phrase mining from text corpora. *Proc. VLDB Endow.* 8(3):305–316.

Erkan, G., and Radev, D. R. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.(JAIR)* 22(1):457–479.

Habernal, I., and Gurevych, I. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics* 43(1):125–179.

Hasan, K. S., and Ng, V. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 1348–1356. Nagoya, Japan: Asian Federation of Natural Language Processing.

Hasan, K. S., and Ng, V. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 751–762. Doha, Qatar: Association for Computational Linguistics.

Lau, J. H.; Newman, D.; and Baldwin, T. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 530–539. Gothenburg, Sweden: Association for Computational Linguistics.

Li, H.; Mukherjee, A.; Si, J.; and Liu, B. 2015. Extracting verb expressions implying negative opinions. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In Marie-Francine Moens, S. S., ed., *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Misra, A.; Anand, P.; Fox Tree, J. E.; and Walker, M. 2015. Using summarization to discover argument facets in online idealogical dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 430–440. Denver, Colorado: Association for Computational Linguistics.

Misra, A.; Oraby, S.; Tandon, S.; TS, S.; Anand, P.; and Walker, M. A. 2017. Summarizing dialogic arguments from social media. In *Proceedings of the 21th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2017)*, 126–136.

Mohammad, S. M.; Sobhani, P.; and Kiritchenko, S. 2017. Stance and sentiment in tweets. *ACM Trans. Internet Technol.* 17(3):26:1–26:23.

Park, J., and Cardie, C. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, 29–38. Baltimore, Maryland: Association for Computational Linguistics.

Paul, M.; Zhai, C.; and Girju, R. 2010. Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 66–76. Cambridge, MA: Association for Computational Linguistics.

Qiu, M., and Jiang, J. 2013. A latent variable model for viewpoint discovery from threaded forum posts. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1031–1040. Atlanta, Georgia: Association for Computational Linguistics.

Stab, C., and Gurevych, I. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 46–56. Doha, Qatar: Association for Computational Linguistics.

Swanson, R.; Ecker, B.; and Walker, M. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 217–226. Prague, Czech Republic: Association for Computational Linguistics.

Thonet, T.; Cabanac, G.; Boughanem, M.; and Pinel-Sauvagnat, K. 2016. *VODUM: A Topic Model Unifying Viewpoint, Topic and Opinion Discovery*. Springer International Publishing. 533–545.

Trabelsi, A., and Zaiane, O. R. 2014. Mining contentious documents using an unsupervised topic model based approach. In *Proceedings of the 2014 IEEE International Conference on Data Mining*, 550–559.

Trabelsi, A., and Zaïane, O. R. 2015. Extraction and clustering of arguing expressions in contentious text. *Data & Knowledge Engineering* 100:226 – 239.

Trabelsi, A., and Zaïane, O. R. 2016. Mining contentious documents. *Knowledge and Information Systems* 48(3):537–560.

Trabelsi, A., and Zaïane, O. R. 2018. Unsupervised model for topic viewpoint discovery in online debates leveraging author interactions. In *Proceedings of the AAAI International Conference on Web and Social Media (ICWSM)*, 425–433. Stanford, California: Association for the Advancement of Artificial Intelligence.

Vilares, D., and He, Y. 2017. Detecting perspectives in political debates. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1573–1582. Copenhagen, Denmark: Association for Computational Linguistics.