# Contrastive Reasons Detection and Clustering from Online Polarized Debates

Amine Trabelsi and Osmar R. Zaïane

Department of Computing Science, University of Alberta
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 on polarized issues. It proposes a pipeline approach centered around the detection and clustering of phrases, assimilated to argument facets using a novel Phrase Author Interaction Topic-Viewpoint model. The evaluation is based on the informativeness, the relevance and the clustering accuracy of extracted reasons. The pipeline approach shows a significant improvement over state-of-the-art methods in contrastive summarization on online debate datasets.

## 1 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 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 [8]. It should make a reader

**Table 1.** Contrastive Digest Table for Abortion.

| *View 1* *Oppose* | | | *View 2* Support | | |
|---|---|---|---|---|---|
| | **Argum. facet label** | **Reason** | | **Argum. 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 |

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 [12]. 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). Note that in many cases the facet is very similar to the reason or proposition initially put forward by a particular viewpoint side, see examples 2 and 6, 7. 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 [9] and [4], 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 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 et al. [14] 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 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. The evaluation of the proposed pipeline is based on three measures: 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 datasets show that our methodology improves significantly over two state-of-the-art methods in terms of documents' summarization, reasons' retrieval and unsupervised contrastive reasons clustering.

## 2 Related Work

The objective of argument mining is to automatically detect the theoretically grounded argumentative structures within the discourse and their relationships [18, 15]. 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 [8] consider rebuttal and refutation as possible components of an argument. Boltužić and Šnajder [3] 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's work [9] on Reason Classification (RC) exploits the reply information associated with the posts. Most of the computational argumentation methods, are supervised. Even the studies focusing on argument identification [19, 13] , 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 [4] 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) [2] 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 [17, 21, 20]. 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 et al. [16] are the first to introduce the problem of contrastive extractive summarization. They applied their general approach 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 [7], is run on scored sentences to generate the summary. Recently, Vilares and He [22] 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.

## 3 Methodology

Our methodology presents a pipeline approach to generate the final digest table of reasons conveyed on a controversial issue. The inputs are raw debate text and the information about the replies. Below we describe the different phases of the pipeline.

### 3.1 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 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. [6], 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 (set according to [6] 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.

### 3.2 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 length, and the information about author-reply interactions in an online debate forum. The objective is to assign a topic and a viewpoint labels 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 length (>=1). Each phrase contain $M_{gda}$ 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 $\theta_{da}$ denote the probability distribution of $K$ topics under a post $d_a$; $\psi_a$ be the probability distributions of $L$ viewpoints for an author $a$; $\phi_{kl}$ be the multinomial probability distribution over words associated with a topic $k$ and a viewpoint $l$; and $\phi_B$ a multinomial distribution of background words. The generative process of a post according to the PhAITV model (see Fig. 1) is the following. An author $a$ chooses a viewpoint $v_{da}$ from the distribution $\psi_a$. For each phrase $g_{da}$ in the post, the author samples a binary route variable $x_{gda}$ from a Bernoulli distribution $\sigma$. It indicates whether the phrase is a topical or a background word. Multi-word phrases cannot belong to the background class. If $x_{gda} = 0$, she samples the word from $\phi_B$. Otherwise, the author, first, draws a topic $z_{gda}$ from $\theta_{da}$, 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 reply is a rebuttal, and $Rb_{id}$ is set to 1 else if it supports, then 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:

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

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

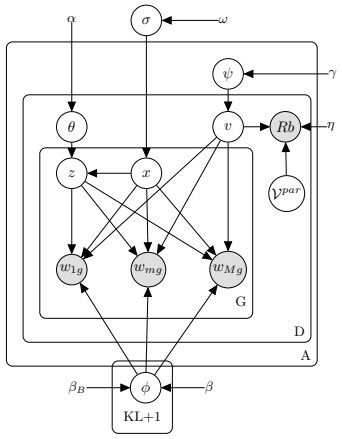


**Fig. 1.** Plate Notation of The PhAITV model

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 Fig. 1, the *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 guiding it by estimating the reply disagreement using methods based on lexicon polarity [17]. 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 (see data statistics in Table 2 as an example). 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 $\boldsymbol{v}_{\neg id}$ is:

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

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:

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

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

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 label $l$. We exploit these assignments to first create clusters $\mathcal{P}_{kl}$s, where each cluster $\mathcal{P}_{kl}$ corresponds to a topic-viewpoint value $kl$. It contains all the phrases that are assigned to $kl$ at least one time. Each phrase $phr$ is associated with its total number of assignments. We note it as $phr.nbAssign$. Second, we rank the phrases inside each cluster according to their assignment frequencies.

### 3.3 Grouping and Facet Labeling

The inputs of this module are Topic-Viewpoint clusters, $\mathcal{P}_{kl}$s, $k = 1..K$, $l = 1..L$, each containing multi-word phrases along with their number of assignments. The outputs are clusters, $\mathcal{A}_l$, of sorted phrases corresponding to argument facet labels for each viewpoint $l$ (see Algorithm 1). 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 [10]. Algorithm 1 proposes a second layer of phrase grouping on each of the constructed Topic-Viewpoint cluster $\mathcal{P}_{kl}$ (lines 3-20). It is based on the number of word 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 (lines 4-7). Then, while iterating on them, we encourage a pair with overlap 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 (lines 8-20).

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 their number of assignments 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 of the contained phrase. If two or more are VE, then the longest phrase, amongst them, adds up the number. Otherwise, we prioritize the most frequently assigned phrase (see lines 21-30 in Algorithm 1). This procedure helps inflate the number of assignments of Verbal Expression phrases in order to promote them to be solid candidates for the argument facet labeling. The final step (lines 32-40) consists of collecting the groups pertaining to each Viewpoint, regardless of the topic, and sorting them based on the cumulative number of assignments of their composing phrases. This will create viewpoint clusters, $\mathcal{C}_l$s, 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 number of assignment to obtain Viewpoint clusters, $\mathcal{A}_l$s, of argument facet labels, sorted in the same order of corresponding groups in $\mathcal{C}_l$s.

### 3.4 Reasons Table Extraction

The inputs of Extraction of Reasons algorithm are sorted facet labels, $\mathcal{A}_l$, for each Viewpoint $l$ (see Algorithm 2). Each label phrase is associated with its sentences $\mathcal{S}_{label}$ where it occurs, and where it is assigned a viewpoint $l$. The target output is the digest table of contrastive reasons $\mathcal{T}$. In order to extract a short sentential reason, given a phrase label, we follow the steps described in Algorithm 2: (1) find, $\mathcal{S}_{label}^{fInters}$, the set of sentences with the most common overlapping words among all the sentences of $\mathcal{S}_{label}$, disregarding the set of words composing the

**Algorithm 1** Grouping and Labeling

**Require:** phrases clusters $\mathcal{P}_{kl}$ for topic $k = 1..K$, view $l = 1..L$
$\mathcal{G}_{kl} \leftarrow \emptyset$ is the set of groups of phrases to create from $\mathcal{P}_{kl}$
**for** each phrase cluster $\mathcal{P}_{kl}$ **do**
  $\mathcal{Q} \leftarrow$ set of all phrase-pairs from phrases in $\mathcal{P}_{kl}$
  **for** each phrase-pair $q$ in $\mathcal{Q}$ **do**
    $q.overlap \leftarrow$ number of word intersections in $q$
  **end for**
  Sort pairs in $\mathcal{Q}$ by number of matches in descending order
  **for** each phrase-pair $q$ in $\mathcal{Q}$ **do**
    **if** $q.overlap \neq 0$ **then**
      **if** $\neg(q.phrase1.grouped) \wedge \neg(q.phrase2.grouped)$ **then**
        New group $grp \leftarrow \{q.phrase1\} \cup \{q.phrase2\}$
        $\mathcal{G}_{kl} \leftarrow \mathcal{G}_{kl} \cup \{grp\}$
      **else if** only one phrase of $q$ in existing $grp'$ **then**
        $grp' \leftarrow grp' \cup \{$non grouped phrase of $q\}$
      **end if**
    **else if** $\neg q.phrase_j.grouped$, $j = 1, 2$ **then**
      New group $grp \leftarrow \{q.phrase_j\}$
      $\mathcal{G}_{kl} \leftarrow \mathcal{G}_{kl} \cup \{grp\}$
    **end if**
  **end for**
  **for** each $grp$ in $\mathcal{G}_{kl}$ **do**
    Sort phrases in $grp$ by giving higher ranking to phrases corresponding to: (1) Verbal Expression; (2) longer phrases; (3) frequently assigned phrases
    **for** each $phr$ in $grp$ **do**
      Find $phr'$ of $grp$ s.t. $phr'.wordSet \subset phr.wordSet$
      **if** $phr'.nbAssign \neq 0$ **then**
        $phr.nbAssign \leftarrow phr.nbAssign + phr'.nbAssign$
        $phr'.nbAssign \leftarrow 0$
      **end if**
    **end for**
  **end for**
**end for**
$\mathcal{C}_l \leftarrow$ set of all groups belonging to any $\mathcal{G}_{*l}$ of view $l$
$\mathcal{A}_l \leftarrow \emptyset$ is the sorted set of all argument facets labels of view $l$
**for** view $l = 1$ to $L$ **do**
  Sort groups in $\mathcal{C}_l$ based on $grp.cumulatifNbAssign$
  **for** each $grp$ in $\mathcal{C}_l$ **do**
    $grp.labelFacet \leftarrow$ phrase with highest $phr.nbAssign$
    $\mathcal{A}_l \leftarrow \mathcal{A}_l \cup \{grp.labelFacet\}$
  **end for**
**end for**
**return** all clusters $\mathcal{A}_l$s of sorted facets' labels for $l = 1..L$

**Table 2.** Datasets Statistics.

| Forum | CreateDebate | | 4Forums | | Reddit |
|---|---|---|---|---|---|
| Dataset | AB | GR | AB | GM | IP |
| # posts | 1876 | 1363 | 7795 | 6782 | 2663 |
| # reason labels | 13 | 9 | - | - | - |
| % arg. sent.[1] | 20.4 | 29.8 | - | - | - |
| % rebuttals | 67.05 | 66.61 | 77.6 | 72.1 | - |

facet label (if the overlap set is empty consider the whole set $\mathcal{S}_{label}$), lines 6-12 in Algorithm 2; (2) choose the shortest sentence amongst $\mathcal{S}_{label}^{fInters}$ (line 13). The process is repeated for all sorted facet labels of $\mathcal{A}_l$ to fill viewpoint column $\mathcal{T}_l$ for $l = 1..L$. Note that duplicate sentences within a viewpoint column are removed. If the same sentence occurs in different columns, we only keep the sentence with the label phrase that has the most number of assignments.Also, we restore stop and rare words of the phrases when rendering them as argument facets. We choose the most frequent sequence in $\mathcal{S}_{label}$.

**Algorithm 2** Extraction of Reasons Digest Table

**Require:** all clusters $\mathcal{A}_l$s of sorted argument facets' labels for view $l = 1..L$;
1: $\mathcal{T}$ is the digest table of contrastive reasons with $\mathcal{T}_l$s columns
2: $\mathcal{T}.columns \leftarrow \emptyset$
3: **for** view $l = 1$ to $L$ **do**
4: $\quad\mathcal{T}_l.cells \leftarrow \emptyset$
5: $\quad$**for** each *label* in $\mathcal{A}_l$ **do**
6: $\quad\quad\mathcal{S}_{label} \leftarrow$ set of all sentences where *label* phrase occurs and assigned view $l$
7: $\quad\quad fInters \leftarrow$ most frequent set of words overlap among $\mathcal{S}_{label}$ s.t. $fInters \neq label.wordSet$
8: $\quad\quad$**if** $fInters \neq \emptyset$ **then**
9: $\quad\quad\quad\mathcal{S}_{label}^{fInters} \leftarrow$ subset of $\mathcal{S}_{label}$ containing $fInters$
10: $\quad\quad$**else**
11: $\quad\quad\quad\mathcal{S}_{label}^{fInters} \leftarrow \mathcal{S}_{label}$
12: $\quad\quad$**end if**
13: $\quad\quad sententialReason \leftarrow$ shortest sentence in $\mathcal{S}_{label}^{fInters}$
14: $\quad\quad\mathcal{T}_l.cells \leftarrow \mathcal{T}_l.cells \cup \{cell(label + sententialReason)\}$
15: $\quad$**end for**
16: $\quad\mathcal{T}.columns \leftarrow \mathcal{T}.columns \cup \{\mathcal{T}_l\}$
17: **end for**
18: **return** $\mathcal{T}$

## 4 Experiments and Results

[1] argumentative sentences in the labeled posts

### 4.1 Datasets

We exploit the reasons corpus constructed by [9] from the online forum CreateDebate.com, and the Internet Argument corpus containing 4Forums.com datasets [1]. We also scraped a Reddit discussion commenting a news article about the March 2018 Gaza clash between Israeli forces and Palestinian protesters [2]. The constructed dataset does not contain any stance labeling. We consider 4 other datasets: Abortion (AB) and Gay Rights (GR) for CreateDebate, and Abortion and Gay Marriage (GM) for 4Forums. Each post in the CreateDebate datasets has a stance label (i.e., support or oppose the issue). The argumentative sentences of the posts have been labeled in [9] with a reason label from a set of predefined reason labels associated with each stance. The reason labels can be assimilated to argument facets or reason types. Only a subset of the posts, for each dataset, has its sentences annotated with reasons. Table 2 presents some statistics about the data. Unlike CreateDebate, 4Forums datasets do not contain any labeling of argumentative sentences or their reasons' types. They contain the ground truth stance labels at the author level. Table 2 reports the percentage of rebuttals as the percentage of replies between authors of opposed stance labels. 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, in CreateDebate, to construct several reference summaries (100) 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 recall measure about the informativeness of the digests, produced on CreateDebate datasets.

### 4.2 Experiments Set Up

We compare the results of our pipeline framework based on **PhAITV** to those of two studies aiming to produce contrastive summarization in any type of contentious text. These correspond to Paul et al.'s [16] and Vilares and He's [22] works. They are based on Topic-Viewpoint models, **TAM**, for the first, and **LAM_LEX** for the second (see Section 2). Below, we refer to the names of the Topic-Viewpoint methods to describe the whole process that is used to produce the final summary or digest. We also compare with a degenerate unigram version of our model that we call **AITV**. AITV's sentences were generated in a similar way to PhAITV's extraction procedure. The difference is that no grouping is involved and the query of retrieval consists of the top three keywords instead of the phrase. As a weak baseline, we generate **random summaries** from the set of possible sentences. We also create **correct summaries** from the subset of reason labeled sentences. One correct summary contains all possible reason types of argumentative sentences for a particular issue. Moreover, we compare

[2] https://www.reddit.com/r/worldnews/comments/8ah8ys/the_us_was_the_only_un_security_council_member_to/

with another degenerate version of our model $\mathbf{PhAITV_{view}}$ which assumes the true values of the posts' viewpoints are given. Note that the objective here is to assess the final output of the framework. Separately evaluating the performance of the Topic Viewpoint model in terms of document clustering has shown satisfiable results. We do not report it here for lack of space.

We try different combinations of the PhAITV's hyperparameters and use the combination which gives a satisfying overall performance. PhAITV hyperparameters are set as follows: $\alpha = 0.1$; $\beta = 1$; $\gamma = 1$; $\beta_B = 0.1$; $\eta = 0.01$; $\omega = 10$; Gibbs Sampling iterations is 1500; number of viewpoints $L$ is 2. We try a different number of topics $K$ for each Topic-Viewpoint model used in the evaluation. The reported results are on the best number of topics found when measuring the Normalized PMI coherence [5] on the Topic-Viewpoint clusters of words. The values of $K$ are 30,10,10 and 50 for PhAITV, LAM_LEX, TAM and AITV, 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 and Israel-Palestine datasets.

### 4.3 Evaluating Argument Facets Detection

The objective is to verify our assumption that the pipeline process, up to the Grouping and Labeling module, produces phrases that can be assimilated to argument facets' labels. We evaluate a total of 60 top distinct phrases produced after 5 runs on Abortion (4Forums) and Gay Rights (CreateDebate). We ask two annotators acquainted with the issues, and familiar with the definition of argument facet (Section 1), to give a score of 0 to a phrase that does not correspond to an argument facet, a score of 1 to a somewhat a facet, and a score of 2 to a clear facet label. Annotator are later asked to find consensus on phrases labeled differently. The average scores, of final annotation, on Abortion and Gay Rights are **1.45** and **1.44**, respectively. The percentages of phrases that are not argument facets are **12.9%** (AB) and **17.4%** (GR). The percentages of clear argument facets labels are **58.06%** (AB) and **62.06%** (GR). These numbers validate our assumption that the pipeline succeeds, to a satisfiable degree, in extracting argument facets labels.

### 4.4 Evaluating Informativeness

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, using annotated reasons of CreateDebate datasets, is explained in Section 4.1. The length of the candidate summaries is proportional to that of the references. Reference summaries on 4Forums or Reddit datasets can not be constructed because no annotation, of reasons and their types, exists. We assess all methods, on CreateDebate, using automatic summary evaluation metric ROUGE [11]. We report the results of Rouge-2's Recall (R-2 R) and F-Measure (R-2 FM). Rouge-2 captures the similarities between sequences of bigrams in references

**Table 3.** Averages of ROUGE Measures (in %, stemming and stop words removal applied) on Abortion and Gay Rights of CreateDebate. Bold denotes best values, notwithstanding Correct Summaries.

| | Abortion | | Gay Rights | |
|---|---|---|---|---|
| | R2-R | R2-FM | R2-R | R2-FM |
| Rand Summ. | 1.0 | 1.0 | 0.7 | 0.8 |
| AITV | 3.0 | 2.8 | **2.7** | **2.8** |
| TAM | 1.8 | 2.1 | 2.0 | 2.4 |
| LAM_LEX | 1.5 | 1.0 | 1.1 | 0.9 |
| PhAITV | **4.5** | **4.6** | **2.7** | **2.8** |
| Correct Summ. | 5.8 | 5.4 | 3.0 | 2.9 |

and candidates. The higher the measure, the better the summary. 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 3, 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 applied.

Table 3 contains the averaged results on 10 generated summaries on Abortion and Gay Rights, respectively. LAM_LEX performs poorly in this task (close to Random summaries) for both datasets. PhAITV performs significantly better than TAM on Abortion, and slightly better on Gay Rights. Moreover, PhAITV shows significant improvement over its degenerate unigram version AITV on Abortion. This shows that phrase modeling and grouping can play a role in extracting more diverse and informative phrases. AITV beats its similar unigram-based summaries on both datasets. This means that the proposed pipeline is effective in terms of summarization even without the phrase modeling. In addition, PhAITV's ROUGE measures on Gay Rights are very similar to those of the correct summaries (Table 3). Examples of the final outputs produced by PhAITV framework and the two contenders on Abortion is presented in Table 5. The example digests produce proportional results to the median results reported in Table 4. 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, and mother's life in danger. However, such informativeness is lacking on both digests of LAM_LEX and TAM. Instead, we remark the recurrence of subjects like killing or taking human life in TAM's digest.

### 4.5 Evaluating Relevance and Clustering

For the following evaluations, we conduct a human annotation task with three annotators. The annotators are acquainted with the studied issues and the pos-

**Table 4.** Median values of Relevance Rate (Rel), NPV and Clustering Accuracy Percentages on CreateDebate, FourForums and Reddit Datasets. Bold denotes best results, notwithstanding PhAITV$_{\text{view}}$.

| | CreateDebate | | | | | | 4Forums | | | | | | Reddit | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Abortion | | | Gay Rights | | | Abortion | | | Gay Marriage | | | Isr/Pal | | |
| | Rel | NPV | Acc. | Rel | NPV | Acc. | Rel | NPV | Acc. | Rel | NPV | Acc. | Rel | NPV | Acc. |
| AITV | 0.66 | 58.33 | 59.09 | 0.5 | **75.0** | 66.66 | 0.66 | 66.66 | 71.42 | 0.5 | 50.0 | 66.66 | 0.6 | 55.55 | 60.00 |
| TAM | 0.53 | 50.00 | 46.42 | 0.5 | 50.0 | 42.85 | 0.33 | 37.50 | 66.66 | 0.3 | 50.0 | 33.33 | 0.3 | 66.66 | 50.00 |
| LAM_LEX | 0.40 | 50.00 | 64.44 | 0.5 | 50.0 | 50.00 | 0.46 | 37.50 | 46.60 | 0.5 | 50.0 | 50.00 | 0.3 | 25.00 | 33.33 |
| PhAITV | **0.93** | **75.00** | **73.62** | **0.8** | **75.0** | **75.00** | **0.80** | **69.44** | **71.79** | **0.7** | **80.0** | **71.42** | **0.9** | **75.00** | **77.77** |
| PhAITV$_{\text{view}}$ | 0.93 | 87.50 | 83.33 | 0.9 | 100 | 100 | 0.80 | 83.33 | 81.81 | 0.9 | 100 | 100 | - | - | - |

**Table 5.** 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.

| 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 [22]** | |
| 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 [16]** | |
| 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? |

sible reasons conveyed by each side. They are given lists of mixed sentences generated by the models. They are asked to indicate the stance of each sentence when it contains any kind of persuasion, reasoning or argumentation from which they could easily infer the stance. Thus, if they label the sentence, the sentence is considered a relevant reason. The average Kappa agreement between the annotators is 0.66. The final annotations correspond to the majority label. In the case of a conflict, we consider the sentence irrelevant.

We consider measuring the relevance by the ratio of the number of relevant sentences divided by the total number of the digest sentences. Table 4 contains the median relevance rate (Rel) over 5 runs of the models, on all datasets. PhAITV-based pipeline realizes very high relevance rates and outperforms its rivals, TAM and LAM_LEX, by a considerable margin on all datasets. Moreover, it beats its unigram counterpart AITV. These results are also showcased in Table 5's examples. The ratio of sentences judged as reasons given to support a stance is higher for PhAITV-based digest. Interestingly, even the PhAITV's sentences judged as irrelevant are not off-topic, and may denote relevant argument facets like "abortion is murder". Results confirm our hypothesis that phrasal facet argument leads to a better reasons' extraction.

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 of pairs of clustered sentences(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 4 contains the median NPV and Accuracy values over 5 runs. Both AITV and PhAITV-based frameworks achieve very encouraging NPV and accuracy results without any supervision. PhAITV outperforms significantly the competing contrastive summarization methods. This confirms the hypothesis that leveraging the reply-interactions, in online debate, helps detect the viewpoints of posts and, hence, correctly cluster the reasons' viewpoints. Table 5 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 4. 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 of PhAITV are not close yet to the $\text{PhAITV}_{\text{view}}$ where the true posts' viewpoint are given. This suggests that the framework can achieve very accurate performances by enhancing viewpoint detection of posts.

## 5 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. The reasons are extracted in a contrastive digest table. 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 issues from online debates show that our PhAITV-based pipeline outperforms state-of-the-art methods for all three criteria. In this research, we dealt with contentious documents in online debate forums, which often 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 the digests suggests the need for improvement in the detection of semantically similar reasons and their hierarchical clustering.

## References

1. Abbott, R., Ecker, B., Anand, P., Walker, M.A.: Internet argument corpus 2.0: An sql schema for dialogic social media and the corpora to go with it. In: LREC (2016)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3**, 993–1022 (2003)
3. Boltužić, F., Šnajder, J.: Back up your stance: Recognizing arguments in online discussions. In: Proceedings of the First Workshop on Argumentation Mining. pp. 49–58. Association for Computational Linguistics, Baltimore, Maryland (June 2014), http://www.aclweb.org/anthology/W14-2107
4. Boltužić, F., Šnajder, J.: Identifying prominent arguments in online debates using semantic textual similarity. In: Proceedings of the 2nd Workshop on Argumentation Mining. pp. 110–115. Association for Computational Linguistics, Denver, CO (June 2015), http://www.aclweb.org/anthology/W15-0514
5. Bouma, G.: Normalized (pointwise) mutual information in collocation extraction. Proceedings of GSCL pp. 31–40 (2009)
6. El-Kishky, A., Song, Y., Wang, C., Voss, C.R., Han, J.: Scalable topical phrase mining from text corpora. Proc. VLDB Endow. **8**(3), 305–316 (Nov 2014). https://doi.org/10.14778/2735508.2735519, http://dx.doi.org/10.14778/2735508.2735519
7. Erkan, G., Radev, D.R.: Lexrank: Graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res.(JAIR) **22**(1), 457–479 (2004)
8. Habernal, I., Gurevych, I.: Argumentation mining in user-generated web discourse. Computational Linguistics **43**(1), 125–179 (2017). https://doi.org/10.1162/COLI_a_00276, https://doi.org/10.1162/COLI_a_00276
9. Hasan, K.S., Ng, V.: 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). pp. 751–762. Association for Computational Linguistics, Doha, Qatar (October 2014), http://www.aclweb.org/anthology/D14-1083
10. Li, H., Mukherjee, A., Si, J., Liu, B.: Extracting verb expressions implying negative opinions. In: Proceedings of the AAAI Conference on Artificial Intelligence (2015), https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9398

11. Lin, C.Y.: 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. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (July 2004)
12. Misra, A., Anand, P., Fox Tree, J.E., Walker, M.: 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. pp. 430–440. Association for Computational Linguistics, Denver, Colorado (May–June 2015), http://www.aclweb.org/anthology/N15-1046
13. Misra, A., Oraby, S., Tandon, S., TS, S., Anand, P., Walker, M.A.: Summarizing dialogic arguments from social media. In: Proceedings of the 21th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2017). pp. 126–136 (August 2017)
14. Mohammad, S.M., Sobhani, P., Kiritchenko, S.: Stance and sentiment in tweets. ACM Trans. Internet Technol. **17**(3), 26:1–26:23 (Jun 2017). https://doi.org/10.1145/3003433, http://doi.acm.org/10.1145/3003433
15. Park, J., Cardie, C.: Identifying appropriate support for propositions in online user comments. In: Proceedings of the First Workshop on Argumentation Mining. pp. 29–38. Association for Computational Linguistics, Baltimore, Maryland (June 2014), http://www.aclweb.org/anthology/W14-2105
16. Paul, M., Zhai, C., Girju, R.: Summarizing contrastive viewpoints in opinionated text. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp. 66–76. Association for Computational Linguistics, Cambridge, MA (October 2010), http://www.aclweb.org/anthology/D10-1007
17. Qiu, M., Jiang, J.: 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. pp. 1031–1040. Association for Computational Linguistics, Atlanta, Georgia (June 2013), http://www.aclweb.org/anthology/N13-1123
18. Stab, C., Gurevych, I.: Identifying argumentative discourse structures in persuasive essays. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 46–56. Association for Computational Linguistics, Doha, Qatar (October 2014), http://www.aclweb.org/anthology/D14-1006
19. Swanson, R., Ecker, B., Walker, M.: Argument mining: Extracting arguments from online dialogue. In: Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue. pp. 217–226. Association for Computational Linguistics, Prague, Czech Republic (September 2015), http://aclweb.org/anthology/W15-4631
20. Thonet, T., Cabanac, G., Boughanem, M., Pinel-Sauvagnat, K.: VODUM: A Topic Model Unifying Viewpoint, Topic and Opinion Discovery, pp. 533–545. Springer International Publishing (2016)
21. Trabelsi, A., Zaiane, O.R.: Mining contentious documents using an unsupervised topic model based approach. In: Proceedings of the 2014 IEEE International Conference on Data Mining. pp. 550–559 (December 2014)
22. Vilares, D., He, Y.: Detecting perspectives in political debates. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 1573–1582. Association for Computational Linguistics, Copenhagen, Denmark (September 2017), https://www.aclweb.org/anthology/D17-1165