Modeling Frames in Argumentation

Yamen Ajjour
Bauhaus-Universität Weimar
Faculty of Media, Webis Group
yamen.ajjour@uni-weimar.de

Milad Alshomary
Paderborn University
Department of Computer Science
milad.alshomary@upb.de

Henning Wachsmuth
Paderborn University
Department of Computer Science
henningw@upb.de

Benno Stein
Bauhaus-Universität Weimar
Faculty of Media, Webis Group
benno.stein@uni-weimar.de

Abstract

In argumentation, framing is used to emphasize a specific aspect of a controversial topic while concealing others. When discussing the legalization of drugs, for instance, its economical aspect may be emphasized. In general, we call a set of arguments that focus on the same aspect a *frame*. An argumentative text has to serve the "right" frame(s) to convince the audience to adopt the author's stance (e.g., being pro or con legalizing drugs). More specifically, an author has to choose frames that fit the audience's interests and cultural background. This paper introduces *frame identification*, which is the task of splitting a set of arguments into a set of non-overlapping frames. We present a fully unsupervised approach to this task, which first removes topical information from the arguments and then identifies frames using clustering. For evaluation purposes, we provide a corpus with 12 326 debate-portal arguments, organized along the frames of the debates' topics. On this corpus, our approach outperforms different strong baselines, achieving an F_1-score of 0.28.

1 Introduction

Different interests, cultural backgrounds, and socializations make people disagree on taking a certain course of action. A *debate* is a means for the involved parties to resolve their disagreement. A debate is characterized by a *topic*, e.g., "Should one legalize abortion?". Upon the topic, the disagreeing parties have a pro or con *stance* respectively, say, "Abortion should be legalized" or "Abortion should not be legalized". A stance is supported by *arguments*; a pro argument could be "Abortion is good for a free society because it gives women their basic right of controlling their bodies." while a con argument could be "Abortion is against human rights because it is a systematic murder of innocent life." One part of an argument (here: the part before the word "because") is called a *conclusion*, the other part consists of one or more *premises*. A debate can be considered as set of pro and con arguments.

Typically, numerous arguments exist for a topic. The parties involved in a debate have to choose among the arguments, thereby *framing* the topic by emphasizing a specific aspect while concealing others. We call the set of the arguments which share an aspect a *frame*. More specifically, a frame F is subset of a set of arguments A, $F \subseteq A$. Likewise, a set of frames, $\{F_1, \ldots, F_k\}$ covers a set of arguments iff. $A \subseteq \bigcup_j^k F_j$.

For instance, the following arguments target different topics but concentrate on the same frame, namely, the *economical* aspect.

Argument 1 *"I support legalization of marijuana since it can be taxed for revenue gain."*
Topic: Marijuana

Argument 2 *"Legalizing prostitution would increase government revenue. A tax on the fee charged by a prostitute, and the imposition of income tax on the earnings of prostitutes would generate revenue."*
Topic: Prostitution

Framing is a decisive step in the construction of arguments, which affects the outcome of a debate (Eemeren and Houtlosser, 1999). To achieve persuasion, an author of an argumentative text should choose frames that resonate with the target audience. As a simple example, an argument appealing to Christianity might not be acceptable to an atheist. Knowing the arguments for a topic along with their frames enables authors to choose those arguments that best address their audience.

The constellation of pro and con arguments for a topic is an urgent need for authors of argumentative texts. Argument search is a new research area that aims at assessing users in forming an opinion and debating. Current approaches use clas-

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 2922–2932,
Hong Kong, China, November 3–7, 2019.

sifiers to mine arguments for a given topic from a relevant document (Levy et al., 2014; Stab et al., 2018; Wachsmuth et al., 2017a). The mentioned approaches ignore identifying the frames of arguments during mining and retrieval, this way omitting extremely valuable information.

The paper in hand starts by reviewing related work to framing (Section 2). In Section 3, we introduce the first argument dataset that has been annotated with frames and topics, and we provide statistical insights into the dataset. Section 4 presents a new unsupervised approach to identify frames in a set of arguments. Our approach first removes topical features from the arguments and then clusters the arguments into frames. In Section 5, we describe the experiments that we conducted to evaluate our approach. We apply the approach to all arguments from *debatepedia.org* and evaluate the returned frames against the ground-truth frames in our dataset. Section 6 shows and discusses the effectiveness of our approach. In Section 7, we analyze the errors made by the approach in the experiments.

The main contributions of this paper are:

- A formal view of frames in argumentation.
- An unsupervised approach to identifying frames in a set of argumentative texts.
- An argument framing dataset with 465 topics, 1 623 frames, and 12 326 arguments.

We freely provide the complete dataset to the research community.[1]

2 Related Work

Research on framing is scattered across different fields such as media, social, and cognitive studies. Entman (1993) was the first to introduce a formal definition of framing as a way to select and make specific aspects of a topic salient. Subsequent research on framing is concentrated on the effect of using frames in news on a specific audience. One of the open questions is whether frames are topic-specific or generic concepts, or both. Vreese (2005) studied framing in news articles and considered frames to be both of the two. Johnson et al. (2017) and Card et al. (2015), on the other hand, defined frames to be independent of the topic and investigated their usage across different topics.

Recently, framing caught some attention in the NLP community. Different computational models have been developed for modeling frames in natural language text. Tsur et al. (2015) used topic models on statements released by congress members of the two major parties in the US, Republicans and Democrats. The learned topics were then aggregated into clusters, such as health and economy, and interpreted as being generic frames. On this basis, the authors studied the frequency of the frames in the released statements for the two parties as well as their distribution over time. A related work was conducted by Menini et al. (2017) to model frames in political manifestos released by the parties (texts declaring a stance) as clusters of key phrases. The developed method was shown to outperform standard topic models in capturing frames.

Card et al. (2015) annotated around 16k news articles on three topics (same-sex marriage, immigration, and smoking), along with a list of 15 generic frames. While the annotations had to cover continuous spans of text, their granularity was left unspecified. The inter-annotator agreement on frames for the different frames ranged between 0.08 and 0.23 in terms of Krippendorff's α. By comparison, our dataset covers both generic and topic-specific frames and are annotated on the argument level. Naderi and Hirst (2017) extended this line of work by training a neural network to classify the frames in the constructed corpus. The authors modeled frames on the sentence level and reached an accuracy of 53.7% in multi-class classification and 89.3% for one-against-others classification. Using the same corpus, (Field et al., 2018) created a lexicon for each one of the 15 frames and analyzed which frames are used mainly to talk about the United States in Russian news.

A related line of research is the mining of arguments from natural language text (Al-Khatib et al., 2016). Most approaches uses supervised classifiers to extract the structure of arguments (*conclusion* and *premise*) (Stab and Gurevych, 2014). (Lawerence and Reed, 2017) showed that topic models helps identifying the relevance of a premise to a conclusion when they are trained on topically relevant documents. The stance of the mined arguments is classified as pro or con towards a given topic (Somasundaran and Wiebe, 2010; Bar-Haim et al., 2017). The arguments are then used for applications such as argument search (Wachsmuth et al.,

[1] Argument framing dataset: https://webis.de/data/webis-argument-framing-19.html or https://doi.org/10.5281/zenodo.3373355

2017a; Levy et al., 2018; Stab et al., 2018) with the goal of retrieving relevant arguments for an input claim. Use cases for argument search include writing and debating support. In comparison to user queries in conventional search that can often be satisfied by one or a few retrieved documents, these use cases require a broader consideration of the retrieved arguments. Hence, the user of an argument search engine will often investigate both stances and multiple frames on a given topic. While several studies tackled the task of ranking arguments according to their quality (Habernal and Gurevych, 2016; Wachsmuth et al., 2017b), how to aggregate arguments into frames is largely unstudied.

The relation between arguments and frames was introduced briefly in some works (Boydstun et al., 2013; Gabrielsen et al., 2011). Still, recent research on computational argumentation largely ignores frames, and a model for aggregating arguments into frames is still missing. Naderi (2016) considered a frame to be an argument and classified sentences in parliamentary speeches into one of seven frames. (Reimers et al., 2019) created a dataset of argument pairs that are labeled according to their similarity. Based on the dataset, they introduced the task of argument clustering which aims at classifying an argument pair with the same topic into similar or dissimilar. The main difference to this work is that no explicit aspect are assigned to the arguments during annotation.

3 Data

Debate portals are websites where people debate or collect arguments for or against controversial topics. Some debate portals are dialogical, such as *debate.org*, allowing two opponents to debate one topic in rounds. Other debate portals such as *debatepedia.org* are wiki-like where arguments are listed according to their stance on the topic. Debate portals keep a canonical structure of the arguments considered for each topic (usually a conclusion and a premise). The structure and the wide topic coverage offered by debate portals has made them a suitable resource for research on computational argumentation (Cabrio and Villata; Al-Khatib et al., 2016; Wachsmuth et al., 2017a).

3.1 Argument Frames from Debatepedia.org

For the given work, we crawled all arguments from debatepedia.org in order to construct a dataset for the evaluation of frame identification. Debatepedia.org organizes a debate into sets of arguments that address a topical aspect of the debate. A label that describes the topical aspect is attached to some of the sets, such as *economics*. An argument on debatepedia.org is listed as a conclusion on the topic along with a premise that supports it.

# Topics	# Frames	# Merged Frames	# Arguments
465	1 645	1 623	12 326

Table 1: Counts of topics, frames, merged frames, and arguments in the webis-argument-framing-19 dataset.

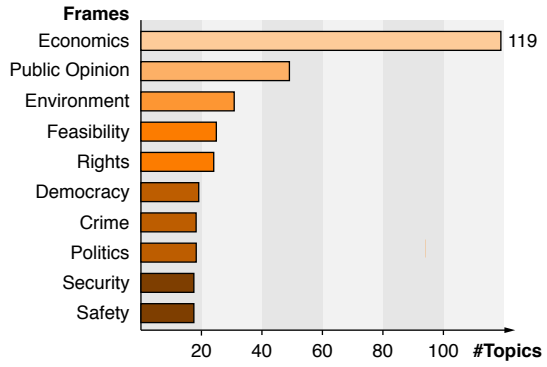

Figure 1: The number of topics in which each of the 10 most frequent frame labels in our dataset occurs.

Arguments which are not labeled might introduce noise to the dataset, since the true knowledge regarding their frames is unavailable. To exclude possible noise in the planned experiments, we filtered out all arguments without labels (about 1800). Next, we analyzed the extracted labels and found that some labels have a similar meaning but are worded differently. In particular, we noticed the presence of the following cases:

1. Labels with hierarchical relations, such as *business* and *US business*.
2. Opposite labels, such as *health* and *unhealthy*, or, *protecting smokers* and *protecting non-smokers*.
3. Labels that are equal when being lemmatized, such as *economics* and *economic*, *democratizing* and *democratic*, etc.

Labels with the same lemmas are likely to carry the same meaning, which is why we merged them into the same label. The count of such merged label pairs was 22, each containing 42 arguments on average. Since the labels in the first and second cases might constitute different frames in some context, we kept them as they are.

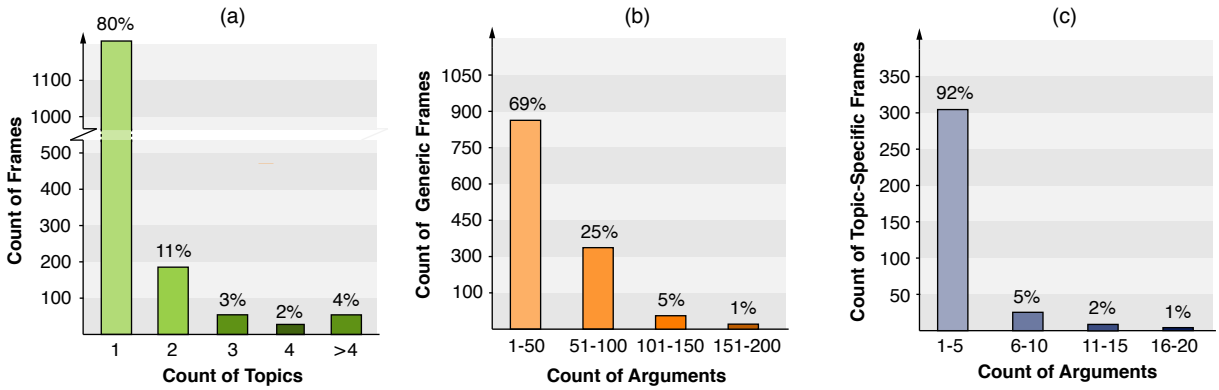

Figure 2: General statistics of frames from debatepedia.org in the webis-argument-framing-19 dataset. (a): Histogram of frames over the count of topics in which they are used. (b): Histogram of generic frames over the count of arguments they contain. (c): Histogram of topic-specific frames over the count of arguments they contain.

3.2 Webis-Argument-Framing-19 Dataset

Table 1 shows general statistics of the final dataset after crawling and preprocessing, called *webis-argument-framing-19*. As visualized in Figure 1, the ten most frequent labels in our dataset are: *economics*, *public opinion*, *environment*, *feasibility*, *rights*, *democracy*, *crime*, *politics*, *security* and, *safety*. These labels largely overlap with those introduced by Card et al. (2015); hence, we considered each set of arguments to be a frame.

The count of topics in which a frame occurs indicates whether a frame is generic or topic-specific. To distinguish between these two types of frames, we grouped all frame labels in our dataset according to how many topics they occur in. Figure 2 (a) shows a histogram of the frames in our dataset over the count of topics in which they are used. As depicted, 80% (1293) of the frames are used in one topic and, hence, we labeled them as *topic-specific*. Frames that are used in more than one topic add up to 20% (330) frames and are labeled as *generic*. Generic frames in the dataset cover 7052 arguments while topic specific frames cover 5274 arguments. Figure 2 (b) and (c) show a histogram of generic and of topic-specific frames over the count of arguments they contain respectively. The histograms reveal that generic frames cover an order of magnitude more as many arguments as topic-specific frames.

4 Approach

In this section, we introduce our unsupervised approach to modeling frames formally. We assume frames to be exclusive and non-overlapping. Given a set of arguments $A = \{a_1, a_2, \ldots, a_n\}$, our goal is to find a set of frames, which constitutes a cover of A. A cover of A is a set of sets $\{F_1, F_2, \ldots, F_k\}$ whose union contains A, i.e., $A \subseteq \bigcup_j^k F_j$. Table 2 lists the symbols used in this section along with their meaning.

The main idea of our approach is to first remove topical features from arguments and then to cluster the arguments into frames. Following known topic modeling approaches, we represent the content of an argument a as a bag of words and propose two models to find topic-specific words. Both models utilize the frequency of the words in an argument and the argument's structure. The structure of a is represented by its conclusion c and its premise(s) p. Our approach includes three main steps:

(a) **Topic clustering.** Cluster the arguments in A into m topics $\bar{\mathcal{A}} = \{\bar{A}_0, \bar{A}_1, \ldots, \bar{A}_m\}$.

(b) **Topic removal.** Given the produced clusters, develop an extraction model E that extracts topical features from an argument a_i and its cluster. E is applied to each $\bar{A}_j \in \bar{\mathcal{A}}$ to remove topic-specific features. As a result, we obtain "topic-free" arguments $a'_i = a_i - E(a, \bar{A}_j)$. We denote the set of all "topic-free" arguments with A' where $A' = \{a'_1, a'_2, \ldots, a'_n\}$.

(c) **Frame clustering.** Cluster the arguments A' into k clusters, each respresenting one frame.

Figure 3 sketches the general idea of the three steps of our approach. We detail our concrete realization of each step in the following.

4.1 Topic Clustering

To cluster the given set of arguments into topics, we first map each argument into a vector space that

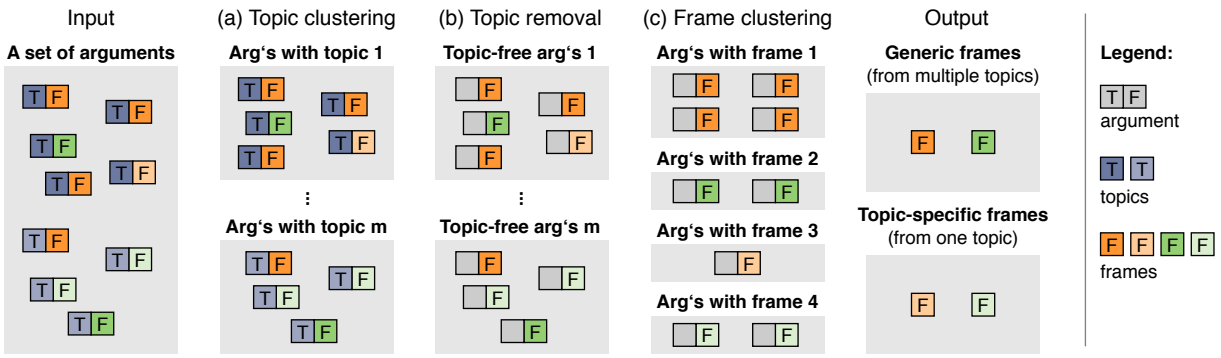

Figure 3: Sketch of the proposed unsupervised approach to argument frame identification. An argument is modeled as a topic and a frame. The input is a sets of arguments. The output is representations of two types of found frames.

Symbol	Meaning
a	An argument
c	The conclusion of an argument
A	A set of arguments
$\bar{A}$	A set of arguments on the same topic
$\mathcal{A}$	A set of sets of arguments
F	A frame
v	A word
V	A vocabulary
E	A topic extraction model

Table 2: Notation of the symbols used in the approach

represents its semantics. We use k-means (Hartigan and Wong, 1979) with Euclidean distance as a clustering algorithm. For semantic spaces, we consider two alternatives: *Term Frequency-Inverse Document Frequency* (TF-IDF), and *Latent Semantic Analysis* (LSA).

TF-IDF TF-IDF defines a vector space whose dimensions are words in the dataset. An argument is mapped to this space according to the frequency of each of its words, normalized by the word's frequency in all considered arguments. TF-IDF is a sparse vector space since all words in a set of arguments are considered. To reduce sparsity, we construct a vocabulary V which comprises the 5000 most frequent words in the arguments after stopword removal. Words that occur in more than half of the arguments are ignored as well. The main reason for reducing the vocabulary is to increase the computational efficiency of the approach.

LSA Latent Semantic analysis (Deerwester et al., 1990) infers from a term-document frequency matrix a linear transformation that project documents into a topic space. We construct two different semantic spaces using LSA. The first, simply called *LSA*, considers each argument to be a document. The second, *LSA Debate*, considers a whole debate to be a document. Since LSA Debate works on the debate level, it can better capture the topic context of an argument. The reason is that arguments capture the topic differently and may have few words in common. Using all arguments in a debate ensures a broader context of the topic. To compare both LSA models systematically, we use the same number of dimension for both models: 1000.

4.2 Topic Removal

The goal of this step is to remove topic-specific features in the topic clusters $\bar{\mathcal{A}} = \{\bar{A}_0, \bar{A}_2, \ldots, \bar{A}_m\}$. To achieve this goal, we develop two models to extract topic-specific features, E_1^q and and E_2. E_1^q utilizes the content of the arguments in one cluster, whereas E_2 utilizes the argument structure, i.e., conclusion and premise information.

E_1^q utilizes the term-frequency inverse document frequency measure TF-IDF for every word v in each cluster. We calculate idf as follows:

$$idf(v) = \frac{\left|\bar{\mathcal{A}}\right|}{\left|\bar{A}_j \in \bar{\mathcal{A}} : v \in \bar{A}_j\right|}$$

Then, $E_1^q(a)$ returns those words that best discriminates a specific topic as follows based on a threshold q, which can be understood as the aggressiveness of the model:[2]

$$E_1^q(a, \bar{A}_j) = \{v \in \bar{A}_j : tf.idf(v) > q\}$$

E_2 utilizes the structure of an argument on a local level. The hypothesis here is that the conclusion of an argument contains more words that target the topic than its premise. Hence, we remove the conclusion in an argument. Formally:

$$E_2(a, \bar{A}_j) = \{v \in c\}$$

[2]In our experiments, we chose the threshold q empirically.

4.3 Frame clustering

This step aims at grouping arguments that share a common aspect after removing topical features. For clustering, we use k-means again and experiment with different values of k. Below, we choose k based on an experiment that evaluates the output of the cluster against the ground-truth. We also use Euclidean distance to estimate the similarity between the arguments in the two semantic spaces.

5 Experiments

Based on the dataset we introduced in Section 3, we conduct experiments to evaluate and analyze our approach to modeling frames in argumentation. As discussed above, the approach consists of three steps: topic clustering, topic removal, and frame clustering. We evaluate the three steps and their interaction with each other in different experiments.

Topic Clustering Experiment The goal of this experiment is to find the best method to group arguments into topics. The produced clusters for each semantic space are evaluated against the arguments' topics in the ground-truth dataset. An external measure is then used to evaluate the output of the clustering algorithm for each semantic space. In particular, we use *Bcubed F_1-score* (Bagga and Baldwin, 1998) to evaluate the effectiveness of our approach in modeling topics in the dataset. Bcubed F_1-score rewards only the instance pairs that exist in the output of the clustering algorithm and in the ground-truth together in the same cluster. The reason for choosing Bcubed F_1-score is that it is proven to satisfy desired constraints in the output of clustering algorithms (Amigó et al., 2009).

Topic Removal Experiment This experiment aims at evaluating our models E_1^q and E_2 at removing topical features from the arguments in $\bar{A}$. The evaluation criterion here is the effectiveness drop of the topic clustering algorithm after removing the topical features in $\bar{A}$. We rerun the topic clustering algorithm with the same k after removing the output of both models E_1^q and E_2. To have a consistent comparison, we set k to the best count of topics we found in the previous experiment.

Frame Clustering Experiment The last experiment evaluates clustering arguments into frames after topic removal. To test our hypothesis that topic removal benefits frame identification, we also cluster arguments in the same semantic space without topic removal. For both semantic spaces, we conduct three experiments: main experiment, generic frame experiment, and topic-specific frame experiment. In the topic-specific and generic frame experiment, we use the frames in our dataset that are labeled as topic specific and generic frame separately. In the main experiment, we test our approach on the whole dataset without distinguishing the type of frames. The different experiments should show us the performance of our approach at identifying generic and topic-specific frames. Similar to topic clustering, we use Bcubed F_1-score (Bagga and Baldwin, 1998) to evaluate the frame clustering algorithm in the three experiment. Since our dataset contains 1623 frames, we evaluate the output of the clustering algorithm for each $k \in \{1000, \ldots, 1600\}$.

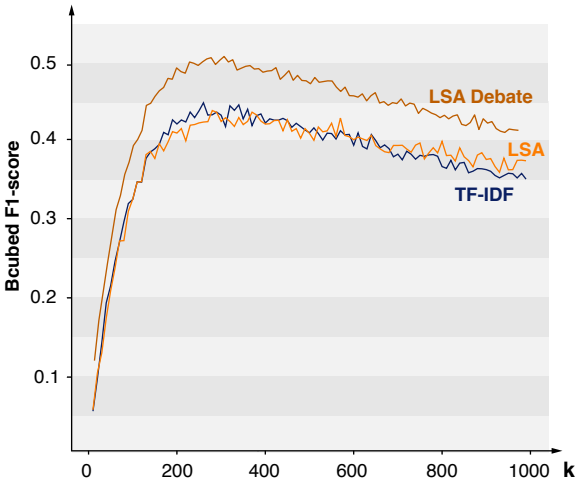

Figure 4: Bcubed F_1-score of the topic clustering algorithms for the semantic spaces TF-IDF, LSA and LSA Debate for each k.

6 Results and Discussion

In the following, we report on the results of the three experiments explained above separately. At the end, we discuss the findings of the experiments and draw final conclusions on the performance of our approach at identifying frames.

6.1 Topic Clustering

Figure 4 shows the effectiveness of topic clustering using the different semantic spaces. We visualize for each k the Bcubed F_1-score of the clustering algorithms for the three semantic spaces. As shown, TF-IDF and LSA perform similarly for all k. The clustering algorithm performs better using the semantic space LSA Debate than LSA and

Semantic Space	# Topics	F_1
LSA Debate	310	**0.52**
TF-IDF	260	0.45
LSA	280	0.44

Table 3: Bcubed F_1-score of the topic clustering algorithm for each semantic space and the corresponding count of topics found.

TF-IDF. This shows the importance of considering the context of an argument for modeling their topics. All the three depicted plots, however, show a clear elbow between topic counts 200 and 400. Table 3 shows the highest corresponding F_1-score and the count of topic clusters for each semantic space. The best topic clustering achieved by the algorithms comprises 310 clusters. Given its high effectiveness in modeling topics, we decided to proceed with the topic clusters produced by the LSA Debate in the next experiment.

6.2 Topic Removal

Table 4 shows the results of the topic removal experiment and frame clustering experiment. For both semantic spaces, the effectiveness of topic clustering algorithm is reported after using the models E_1^q and E_2 to remove topic-specific words. To evaluate the topic extraction models, we re-list the effectiveness achieved by the topic clustering algorithm for both spaces. We show the results of E_1^q only for $q = 0.005$ since higher values of q showed similar or lower results in all experiments.

As shown, E_2 decreases the effectiveness of topic clustering algorithm to around the half. The model $E_1^{0.005}$ achieves a smaller drop of 0.03-0.04 in the two semantic spaces. Despite its simplicity, E_2 is more effective at removing topic-specific features than $E_1^{0.005}$.

6.3 Frame Clustering

Table 4 shows the results of the frame clustering algorithm in the experiments: generic, topic-specific, and main. In each experiment, the clustering algorithm is run after using the two topic extraction models to remove topic-specific features and without applying them (baseline). In the main and the generic experiment, using the topic extraction models outperforms not using them in both semantic spaces. In the topic-specific experiment, our approach's effectiveness outperformed the baseline only in LSA space. The comparison between the results in the generic and topic-specific experiments shows that identifying generic frames is harder. The reason can also be the small size of topic-specific frames in the ground-truth. Our approach, however, is only effective at identifying generic frames and fails at outperforming the baseline in the topic-specific experiments. A reason to justify this is that removing topic-specific features negatively affects identifying topic-specific frames.

To better analyze our approach, we plot the achieved Bcubed F_1-score for each semantic space and each experiment for different values of k. Figure 5 (a,b,c) shows the effectiveness achieved in the three experiments main, topic-specific and generic respectively in TF-IDF space. As shown, both models $E_1^{0.005}$ and E_2 start to outperform the baseline at $k = 1200$ in the main experiment. All the approaches converge starting from this value and not much effectiveness is achieved for higher values of k. In the generic experiment, both models achieve their first peaks at $k = 800$. The performance of both models oscillates but keeps at the same rate of for larger values of k. In the topic-specific experiments, the performance of our approach increases significantly while approaching the value of $k = 400$. The gain achieved by for more clusters decreases slowly for larger values of k.

Figure 6 (a,b,c) shows the effectiveness achieved in the three experiments main, topic-specific and generic respectively in LSA space. As depicted in the three figures, the model $E_1^{0.005}$ outperforms E_2 in all cases which shows that content-based topic-removal of arguments is more effective than using its structure. In the generic experiment, all models in LSA space shows subpar effectiveness to their counterpart in TF-IDF space and lack clear peaks. In the topic-specific experiment, our approach outperforms the LSA baseline and their counterpart in TF-IDF space. Nevertheless, like in the generic experiment, no clear peak is reached by any model.

6.4 Discussion

The results show the merit of removing the topic-specific of an argument for identifying its frame. According to the reported results, our approach is effective at identifying generic frames and does not suit identifying topic-specific frames. An interesting finding is that the premise of an argument carries more information about its frame than the conclusion. This is shown in the higher effectiveness achieved after applying E_2 compared to the

Semantic Space	Topic Removal		Frame Clustering								
	Model	Topic F_1	Generic Frames			Topic-specific Frames			Frames		
			F_1	P	R	F_1	P	R	F_1	P	R
TF-IDF	Baseline	**0.45**	0.19	0.25	0.15	**0.48**	0.53	0.44	0.26	0.27	0.25
	$E_1^{0.005}$	0.42	**0.28**	0.26	0.30	0.45	0.50	0.40	**0.28**	0.24	0.33
	E^2	0.17	0.26	0.25	0.28	0.45	0.48	0.42	0.27	0.25	0.29
LSA	Baseline	0.44	0.16	0.20	0.13	0.39	0.44	0.35	0.21	0.22	0.22
	$E_1^{0.005}$	0.4	0.21	0.15	0.33	0.47	0.44	0.48	0.26	0.25	0.27
	E^2	0.25	0.2	0.18	0.22	0.46	0.41	0.50	0.24	0.24	0.24

Table 4: Best bcubed F_1-score, precision, and recall for the topic extraction models E_1^q, E_2 and without topic removal (baseline) in the generic, topic-specific and main frame experiments together with the corresponding bcubed F_1-score in topic clustering.

baseline. A justification can be that a conclusion is more likely to carry stance-taking words toward the topic. In general, $E_1^{0.005}$ achieved higher results than E_2, which shows that using the content of an argument is more effective than using its structure to model frames. A possible justification for this can be that E_2 is more aggressive than needed at removing topic-specific features.

7 Error Analysis

We analyze the topic and frame clusters produced by our approach to convey to the reader a sense of its performance. For topic clusters, we focus on the semantic space LSA Debate since our approach performed the best in this semantic space. For Frame clusters, we analyze the output of our approach in the semantic space TF-IDF after applying $E_2^{0.005}$ since our approach performed the best in this semantic space. Our goal is to identify the topics and frames in the dataset which our approach completely confused or correctly identified. To identify these cases, we sort the topics and frames in the ground-truth dataset according to the maximum F_1-score achieved in the aforementioned semantic spaces respectively. We manually analyze the topic and frames labels and the count of arguments they comprise and report the most interesting cases.

For topic clustering, examples of topics that our approach correctly identified (with an F_1-score of 1) are: *Zoos* and *Compulsory vaccination*. On the other hand, our approach struggled at identifying topics like *Is Pluto a planet?* and *Immunity from prosecution for politicians* (with an F_1-score lower than 0.1). A reason for this might be that these topics are too specific and not covered well in our dataset.

In frame clustering, the hardest cases for our approach in TF-IDF space were topic-specific frames that contain few arguments, e.g., *child disability*. Generic frames such as *rights* and *feasibility* were also hard to identify (with an F_1-score lower than 0.1). A possible explanation is that these frames can be confused with generic frames like *human rights* and *economics*. Examples of generic frames that were effectively identified are *freedom of speech* and *public health* (with an F_1-score equals to 0.5).

8 Conclusion

A disagreement between people on a topic can lead to a lively debate where the opposing parties exchange arguments on the topic to enforce their stance. In favor of a particular stance, an argument emphasizes a certain aspect of the topic, thereby hiding other aspects. This phenomenon is called framing and has been introduced in social science (Entman, 1993). Knowing the frame of an argument helps users to choose arguments that better address the audience's cultural background.

Research on framing in natural languages is still lacking. In this paper, we tried to close this gap by introducing a formal view on framing that defines a frame as a set of arguments that share an aspect. Starting from this view, we introduced an approach to remove the topic's features from the arguments and then to cluster them. We operationalized our approach by using two different models. While the first removes an argument's topic features using its content, the second utilizes its structure (conclusion and premise).

For evaluation purposes, we constructed a new dataset that comprises 12 326 arguments grouped along 1 623 frames based on debatepedia.org. The experiments show that we can outperform sensi-

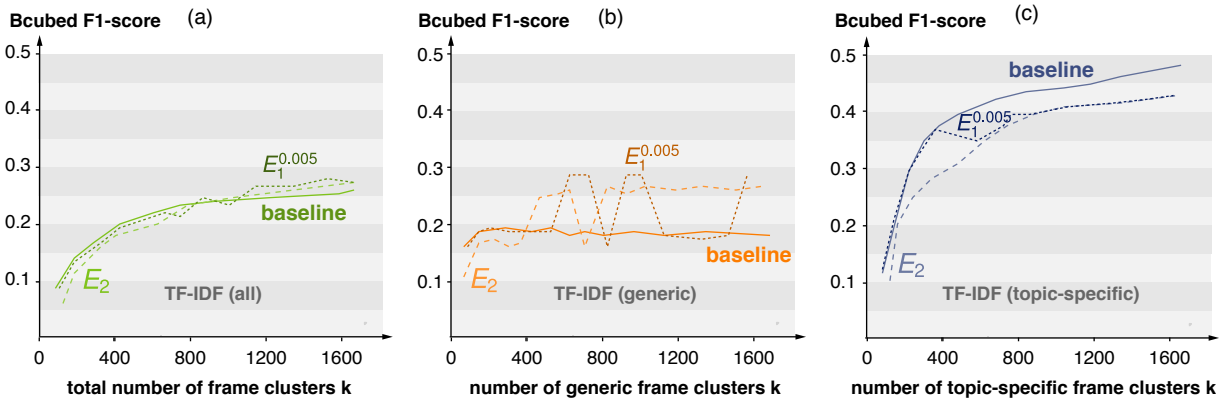

Figure 5: Effectiveness of frame clustering with TF-IDF without topic removal (baseline) and after applying $E_1^{0.005}$ and E_2 in (a) the main frame experiment, (b) the generic frame experiment, and (c) the topic-specific experiment.

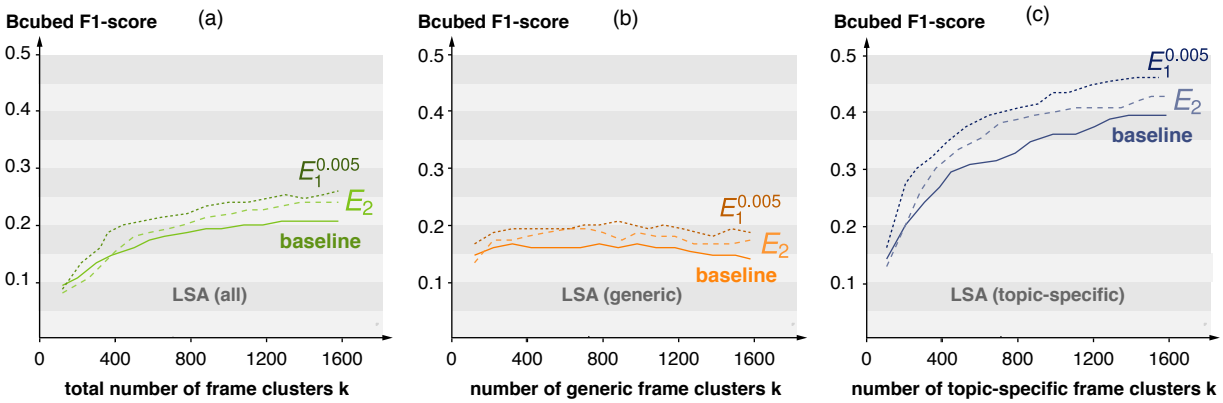

Figure 6: Effectiveness of frame clustering with LSA without topic removal (baseline) and after applying $E_1^{0.005}$ and E_2 in (a) the main frame experiment, (b) the generic frame experiment, and (c) the topic-specific experiment.

ble baselines that utilize the same semantic spaces without deleting topic information. We conducted three types of experiments that evaluate our approach in identifying generic (those which are used in multiple topics), topic-specific and both.

Our experiments clearly show the benefit of removing topic's features for identifying an argument's frame. In particular, we find that identifying generic frames benefits from removing topic features, which are actually the hardest case. On the other hand, removing topic features cannot help in identifying topic-specific frames. We also observed that removing an argument's conclusion helps identifying its frame, although it is more likely to carry the stance and the topic of an argument.

Having set a lacking methodology for modeling frames in argumentation, our next step is to develop better approaches for modeling and removing the topic of an argument. Neural networks such as auto-encoders and attention-based models are likely to perform better at modeling frames in argumentation than LSA and TF-IDF. Regarding topic removal, potential research directions will investigate using external knowledge such as Wikipedia to find topic-specific features.

Future work on framing will focus on its application to down-stream argument mining tasks such as analyzing argument quality. Especially interesting is whether specific frames are expected to persuade an audience. A follow-up question will be whether frames in an argumentative text should be delivered in a specific sequence to achieve the persuasion of an audience.

The simplicity of our model allows its application in domains such as news, laws, or student essays. A promising application of our model is argument search since a frame of an argument sheds a light on its acceptability for a specific audience. A clear problem here is how to label a frame given its arguments in order to deliver short labels for the user. We also expect framing to play a major role in generating arguments since a specific frame might resonate with an audience.

References

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016. Cross-Domain Mining of Argumentative Text through Distant Supervision. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2016)*, pages 1395–1404. Association for Computational Linguistics.

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Fleisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. 12:461–486.

Amit Bagga and Breck Baldwin. 1998. Algorithms for Scoring Coreference Chains. In *The 1st International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance Classification of Context-Dependent Claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 251–261. Association for Computational Linguistics.

Amber E. Boydstun, Dallas Card, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2013. Identifying media frames and frame dynamics within and across policy issues. In *Proceedings of the Workshop on New Directions in Analyzing Text as Data*.

Elena Cabrio and Serena Villata. Natural Language Arguments: A Combined Approach. In *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI 12)*, pages 205–210, Amsterdam, The Netherlands. IOS Press.

Dallas Card, Amber E Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The Media Frames Corpus: Annotations of Frames Across Issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL 2015)*, pages 438–444. Association for Computational Linguistics.

Scott Deerwester, Susan T. Dumals, George W. Furnasand, Thomas K Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. 41(6):391–407.

Frans Van Eemeren and Peter Houtlosser. 1999. Strategic Manoeuvring in Argumentative Discourse. *Discourse Studies*, 1(4):479–497.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Anjalie Field, Doron Kliger, Shuly Wintner, Jinnifer Pan, Dan Jurafsky, and Yulia Tesvetkov. 2018. Framing and Agenda-setting in Russian News: a Computational Analysis of Intricate Political Strategies.

Jonas Gabrielsen, Sine NØrholm Just, and Mette Bengtsson. 2011. Concepts and Contexts – Argumentative Forms of Framing. *Proceedings of the 7th Conference of the International Society for the Study of Argumentation*, pages 533–543.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? Analyzing and predicting convincingnessof Web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 1589–1599. Association for Computational Linguistics.

J. A. Hartigan and M. A. Wong. 1979. A k-means clustering algorithm. *JSTOR: Applied Statistics*, 28:100–108.

Kirsten Johnson, Di Jin, and Dan Goldwasser. 2017. Modeling of Political Discourse Framing on Twitter. In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM 2017)*, pages 556–559. Association for the Advancement of Artificial Intelligence.

John Lawerence and Chris Reed. 2017. Mining Argumentative Structure from natural language text using Automatically Generated Premise-Conclusion Topic Models. In *Proceedings of the 4th Workshop on Argumant Mining (ArgMining 2017)*, pages 118–128. Association for Computational Linguistics.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context Dependent Claim Detection. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers (COLING 2015)*, pages 1489–1500. Dublin City University and Association for Computational Linguistics.

Ran Levy, Ben Boginand Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018. Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 2066–2081.

Stefano Menini, Federico Nanni, Simone Paolo Ponzetto, and Sara Tonelli. 2017. Topic-Based Agreement and Disagreement in US Electoral Manifestos. In *Proceedings of the 2017 Conference of Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 2938–2944.

Nona Naderi. 2016. Argumentation Mining in Parlimentary Discourse. In *Proceedings of the 11th International Conference of the Ontario Society for the Study of Argumentation*, pages 1–9.

Nona Naderi and Graeme Hirst. 2017. Classifying Frames at the Sentence Level in News Articles. In

Proceedings of Recent Advances in Natural Language Processing 2017 (RANLP 2017), pages 536–542.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and Clustering of Arguments with Contextualized Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 567–578. Association for Computational Linguistics.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing Stances in Ideological On-Line Debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124.

Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. ArgumenText: Searching for Arguments in Heterogeneous Sources. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: System Demonstrations (NAACL 2018)*, pages 21–25. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. Identifying Argumentative Discourse Structures in Persuasive Essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 46–56. Association for Computational Linguistics.

Oren Tsur, Dan Calacci, and David Lazer. 2015. A Frame of Mind: Using Statistical Models for Detection of Framing and Agenda Setting Campaigns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*, pages 1629–1638.

Claes H. De Vreese. 2005. News Framing: Theory and typology. *Information Design Journal and Document Design*, 13(1).

Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017a. Building an Argument Search Engine for the Web. In *Proceedings of the Fourth Workshop on Argument Mining (ArgMining 2017)*, pages 49–59. Association for Computational Linguistics.

Henning Wachsmuth, Benno Stein, and Yamen Ajjour. 2017b. "PageRank" for Argument Relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, pages 1116–1126. Association for Computational Linguistics.