Automatic Identification and Naming of Overlapping and Topic-specific Argumentation Frames

Carolin Schindler[1], Annalena Aicher[1], Niklas Rach[2], Wolfgang Minker[1],

[1]Institute of Communications Engineering, Ulm University, Germany
[2]Tensor AI Solutions GmbH, Germany
Correspondence: carolin.schindler@uni-ulm.de

Abstract

Being aware of frames, i. e., the aspect-based grouping of arguments, is crucial in applications that build upon a corpus of arguments, allowing, among others, biases and filter bubbles to be mitigated. However, manually identifying and naming these frames can be time-consuming and therefore not feasible for larger datasets. Within this work, we present a sequential three-step pipeline for automating this task in a data-driven manner. After embedding the arguments, we apply clustering algorithms for identifying the frames and subsequently, utilize methods from the field of cluster labeling to name the frames. The proposed approach is tailored towards the requirements of practical applications where arguments may not be easily split into their argumentative units and hence can belong to more than one frame. Performing a component-wise evaluation, we determine the best-performing configuration of the pipeline. Our results indicate that frames should be identified by performing overlapping and not exclusive clustering and the naming of frames can be accomplished best by extracting aspect terms and weighting them with c-TF-IDF.

1 Introduction

By "select[ing] some aspects of a perceived reality and mak[ing] them more salient in a communicating text" (Entman, 1993, p. 52), framing introduces a bias in the presentation of information. Hence, applications utilizing among others argument mining (Skiera et al., 2022), argument search (Ajjour et al., 2019), discourse analysis (Ruckdeschel and Wiedemann, 2022), summarization (Misra et al., 2016), or argumentative dialogue (Rach et al., 2018; Aicher et al., 2019) need to be aware of the frames that are present in their data.

Within this work, we present a pipeline for automatically identifying and naming such topic-specific frames among a collection of arguments. Thereby, we consider the overlapping nature of the task, i. e., that an argument can belong to more than one frame (Reimers et al., 2019; Dumani et al., 2021; Ruckdeschel and Wiedemann, 2022).

Current works identifying argumentation frames apply an exclusive mapping of arguments to frames and leave the naming of the identified frames to future work. While Reimers et al. (2019) and Daxenberger et al. (2020) directly state these limitations as directions for future work, Dumani et al. (2021) justify the exclusive clustering procedure by assuming that the arguments are provided in elementary parts, i. e. argument units (Trautmann et al., 2020), that belong to exactly one frame. However, this assumption is not always viable for a practical application to arguments "in the wild" since argument unit extraction itself is not an easy task (Stab et al., 2018; Trautmann et al., 2020). Therefore, we focus on creating an overlapping clustering for identifying frames and do not exclude their naming from the task.

Following the conceptual discussion in Schindler (2024), we perform the automatic identification and naming of frames in three sequential steps. First, the arguments need to be embedded in an embedding space that is capable of capturing aspect-based similarity. With this notion of similarity, we then cluster the arguments into frames, thereby considering the overlapping nature of the task. Afterwards, we utilize methods from the field of cluster labeling to name the identified frames. In a component-wise evaluation setup, we identify the best performing approach for each step. In the course of this, we demonstrate that the identification of frames benefits from applying overlapping clustering algorithms on the argument-level and show that the naming performs best when building upon aspect-based candidate extraction.

The remainder is organized as follows: After clarifying the terminology used throughout this work in Section 2, Section 3 gives an overview over

Proceedings of the 12th Argument Mining Workshop, pages 147–159
July 31, 2025 ©2025 Association for Computational Linguistics

related work. We detail the individual steps of our pipeline for identifying and naming overlapping, topic-specific argumentation frames in Section 4. The different approaches of performing these steps are evaluated in Section 5 along with a discussion of the results before we conclude in Section 6.

2 Terminology

Following the definition of arguments by Stab and Gurevych (2014), in this work, an *argument* is a sentence that is making a defeasible point and is having a stance towards a debatable topic. Such an argument is built from one or more *argument units* (Trautmann et al., 2020), i. e. indivisible argumentative spans that can be used in different combination in other arguments as well. The *aspects* of an argument "hold the core reason upon which the conclusion/evidence is built" (Schiller et al., 2021, p. 380). The tokens of an argument that are indicative for the aspect(s) addressed by it are *aspect terms* (Trautmann, 2020). When arguments that are addressing similar aspects of the topic are grouped together, the resulting group is a *frame*. In this work, frames are topic-specific, independent of the stance of the arguments, non-redundant, and can be named succinctly in a human-understandable manner by a *frame label*. The grouping of arguments into frames can be also viewed as a grouping of similar aspects into *aspect categories*. Since an argument can address multiple aspects of the topic that not necessarily need to be grouped into the same aspect category, an argumentative sentence can belong to more than one frame (Reimers et al., 2019; Dumani et al., 2021; Ruckdeschel and Wiedemann, 2022).

3 Related Work

Without automation, the identification and naming of frames needs to be performed manually in a time consuming process for every topic individually (Lai et al., 2022; Jurkschat et al., 2022; Ruckdeschel and Wiedemann, 2022). The topic-independent automation, however, is a challenging task. Yet, it is little known about what features are relevant for grouping arguments with respect to the aspects they address (Opitz et al., 2021), but fine-grained semantic nuance might already be crucial (Reimers et al., 2019). Further, there is no general guideline for creating or naming frames, leaving room for subjectivity in the process (Lai et al., 2022; Jurkschat et al., 2022; Ruckdeschel and Wiedemann, 2022).

Identification of Frames When identifying frames with the help of clustering, one performs aspect-based argument clustering. For aspect-based argument clustering, there are no frame labels given and labeling the resulting clusters is often left to future work. By formulating the clustering problem as a similarity scoring task between pairs of arguments, one can perform supervised training; either by regression with a graded scale (Misra et al., 2016) or by classification with a binary labeling scheme (Reimers et al., 2019). Even with little training data, the supervised approach outperforms the unsupervised methods in a cross-topic evaluation setup (Reimers et al., 2019). Moreover, Reimers et al. (2019) point out that exclusive clustering algorithms are a sub-optimal choice since they do not reflect the properties of the data: In 21.9% of the cases the transitivity property induced by exclusive clustering is violated in their dataset. Hence, the overlapping nature of aspect-based argument clustering should be taken into account as in the herein presented work.

Operating on the term-level with an exclusive clustering approach as in Lai et al. (2022) for aspect-based document clustering, was additionally inspired by the following works. Ruckdeschel and Wiedemann (2022) performed an investigation on the level at which frames should be coded. For annotating arguments with a predefined set of frame labels, they found the token-level to be best-suited. In a multi-label argument classification setting, their results suggest that it is beneficial to consider a more granular level than sentence-level. An unsupervised clustering approach proposed by Heinisch and Cimiano (2021) groups fine-granular, topic-specific aspects into more general aspect categories. There, no names are derived for the created categories.

Naming of Frames IBM Project Debater (Slonim et al., 2021; Bar-Haim et al., 2021) makes use of Wikipedia titles that are related to the individual argumentative sentences in order to exclusively cluster and label them. In the summetix API[1], formerly known as ArgumenText API (Daxenberger et al., 2020), a labeling of the clusters is implemented (Skiera et al., 2022) on top of an exclusive clustering with fine-tuned embeddings (Reimers et al., 2019). The label of each cluster is the aspect term with the highest c-TF-IDF (Grootendorst, 2022) score. This

[1] https://api.summetix.com/

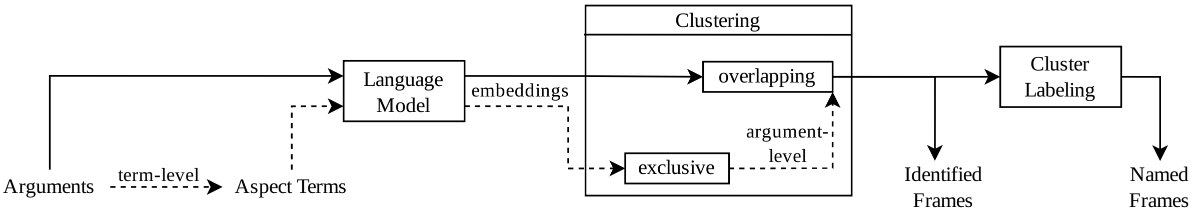

Figure 1: Pipeline for automatic identification and naming of argumentation frames.

is one of the approaches that we are going to test for the naming of overlapping frames.

Topic Modelling Given a collection of texts, *topic modelling* (Churchill and Singh, 2022) aims at structuring these texts by identifying and naming the topics they belong to. Topic modelling is commonly performed on a set of documents, however, it can also be applied to a set of arguments that are belonging to different but unknown topics (Ajjour et al., 2019; Färber and Steyer, 2021). Hence, topic modelling is similar to our task but operating on topics instead of frames. The neural topic model BERTopic (Grootendorst, 2022) is noteworthy in the scope of this work since our steps for identifying and naming frames are similar to theirs for topic modelling and Haddadan et al. (2022) already have applied BERTopic for the qualitative analysis of a dataset into frames. However, our contribution goes beyond the pipeline approach by considering and investigating the overlapping nature of the resulting clustering and providing a quantitative evaluation.

Related Tasks Formally, *frame detection* is a supervised multi-label argument classification task with a predefined set of frame labels (Mou et al., 2022). We, however, do not have any frame labels given and hence we would need to obtain them in a data-driven manner first before performing zero-shot frame detection (Syed et al., 2023; Mou et al., 2022; Ajjour et al., 2018). Moreover, the generalization of classifiers to unseen topics and label sets poses a major challenge for zero-shot approaches. Given a set of arguments, *key point analysis* (Bar-Haim et al., 2020a,b; Friedman et al., 2021) aims at creating a list of prominent key points and then matching the arguments to these key points. This results in every key point being the label for an individual group of arguments. Differently to our task, the groups are stance-dependent and the key points labelling the clusters are argumentative sentences that could be utilized for creating a summary on the topic.

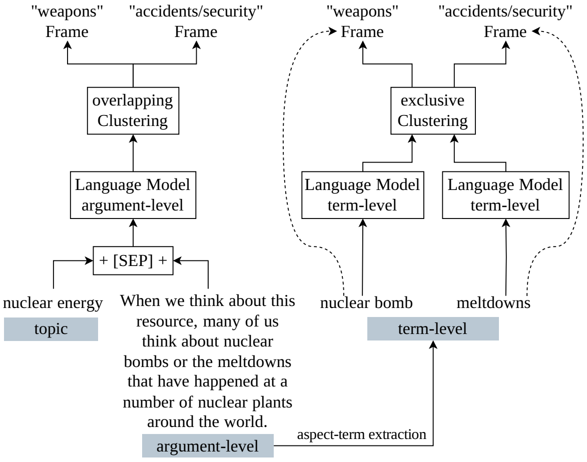
"weapons" Frame
"accidents/security" Frame
overlapping Clustering
Language Model argument-level
+ [SEP] +
nuclear energy
topic
When we think about this resource, many of us think about nuclear bombs or the meltdowns that have happened at a number of nuclear plants around the world.
argument-level
aspect-term extraction
"weapons" Frame
"accidents/security" Frame
exclusive Clustering
Language Model term-level
Language Model term-level
nuclear bomb
meltdowns
term-level

Figure 2: Exemplary processing of an argument from the AAC dataset in our pipeline.

4 Pipeline for Automatic Identification and Naming of Frames

Our pipeline for identifying and naming frames is depicted in Figure 1 and entails three sequential steps, which are described in more detail below. Given a set of n argumentative sentences $\mathcal{S} = \{s_1, ..., s_n\}$ about the topic t, we first require embeddings that capture the notion of aspect-based similarity. Second, we perform a complete clustering $\mathcal{C} = \{c_1, ..., c_k\}$ of the embeddings into k distinct frames with k not being known in advance. Finally, we apply methods from the field of cluster labeling to name the frames with a frame label.

Besides performing the clustering in an overlapping manner on the argument-level, we also experiment with applying exclusive clustering algorithms on the term-level and mapping the resulting clustering back to the argument-level. On the argument-level, the sentential argument itself is treated as the unit to be clustered, while on the term-level we extract the aspect terms from the arguments and utilize these for further processing. An example of performing the frame identification on the argument- and term-level is provided in Figure 2.

4.1 Embeddings for Aspect-based Similarity

To gain embeddings that are suited for aspect-based similarity, it is infeasible to utilize pre-trained language models since they are tailored towards the notion of semantic textual similarity (STS) (Cer et al., 2017). Whereas STS measures the similarity of two texts based on their meaning, we are concerned with the similarity of the aspects addressed by the arguments irrespective of the actual point they are making regarding the aspect (Misra et al., 2016). Hence, when the task involves aspect-based properties, unsupervised models that are pre-trained on semantic properties are outperformed by their fine-tuned counterparts with supervision (Reimers et al., 2019; Dumani et al., 2021).

A task that is utilized for fine-tuning embeddings for aspect-based similarity is aspect-based similarity prediction, where the model has to decide whether two arguments s_1 and s_2 about the same topic t are similar in terms of the aspects they address. In the course of this, every argument is embedded by the model individually and the cosine similarity between the embeddings serves as a measure for aspect-based similarity. The binary classification decision is made by applying a threshold to the predicted similarity score. During fine-tuning, there is also the option to not use a binary but a graded label set, reflecting the circumstance that aspect-based similarity is not a discrete decision (Misra et al., 2016).

In Schindler (2024), the STS-based embeddings of the SBERT (Reimers and Gurevych, 2019) model *all-mpnet-base-v2* were fine-tuned. Following the experimental procedure of Reimers et al. (2019), they performed a four-fold cross-topic validation on the Argument Aspect Similarity (UKP ASPECT) Corpus[2] (Reimers et al., 2019). Depending on the level of granularity that they were operating on, they tested different kinds of information for creating the embeddings.

Here, we employ their respective best performing model in the pipeline. On the argument-level, we utilize their SBERT model fine-tuned on all topics of the UKP ASPECT corpus with the topic t prepended to the argumentative sentence s as an input. With this input configuration, they achieved human-like performance in the four-fold cross-topic validation. On the term-level, the input to the model is a single aspect term AT and no fine-tuning is performed. The aspect terms are extracted from the arguments by querying the summetix API[3] (Schiller et al., 2021).

Their results are confirming that STS is not a good indicator for aspect-based similarity on the argument-level. Nevertheless, STS is performing well on the level of aspect terms. This observation can be explained by the fact that the meaning of terms grouped into an aspect category should be similar. However, for an argumentative sentence, meaning is evaluated on a larger scale than aspects.

4.2 Identification of Frames

We make use of clustering to identify frames, i. e., group the arguments by the aspects they address. The cosine distance between the embeddings of the items serves as the distance measure. To account for the curse of dimensionality that distance measures are prone to (Aggarwal et al., 2001; Steinbach et al., 2004), we apply dimensionality reduction on a per topic basis as a preprocessing step.

For a comparison of overlapping clustering on the argument-level and exclusive clustering on the term-level, which is afterwards mapped back to the argument-level, we select the following equivalent centroid-based clustering algorithms: k-means (MacQueen, 1967) as an exclusive clustering algorithm, which has already been applied for similar frame identification tasks in previous works (e. g., Färber and Steyer (2021); Ajjour et al. (2019); Heinisch and Cimiano (2021)), and fuzzy c-mean (FCM) (Bezdek et al., 1984) as a soft/fuzzy clustering algorithm whose output can be transformed into a hard overlapping clustering. We approach this transformation by assigning every clustered item to the clusters with the highest scores until the cumulative sum of cluster scores that the item is assigned to exceed the threshold θ_{cum}. If multiple clusters are having the same score for an item, we select all of them simultaneously.

4.3 Naming of Frames

For automatically naming the identified frames, we apply a differential cluster labeling strategy that is agnostic of the other frames. This way, the name of the frame label of each frame solely depends on the arguments within the frame and the complete collection of arguments.

First, we generate a set of candidates for each frame. The candidates are either lemmatized as-

[2] https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/1998

[3] https://api.summetix.com/

pect terms *ATs* extracted from the arguments as in Skiera et al. (2022) or terms *FTs* extracted from the lemmatized collection of arguments without stop words based on Luhn's expressiveness of terms assumption (Luhn, 1958). This assumption states that the most important terms are those with mid frequencies, i. e., that neither occur too frequently nor too rarely. To this end, we consider the following approaches, where *frame* and *¬frame* denote that the terms are selected within the frame or among the arguments outside the frame, respectively:

(A) FTs_{frame}

(B) ATs_{frame}

(C) $FTs_{frame} \setminus FTs_{\neg frame}$

(D) $ATs_{frame} \setminus ATs_{\neg frame}$

(E) $FTs_{frame} \cap ATs_{frame}$

(F) $(FTs_{frame} \setminus FTs_{\neg frame}) \cap (ATs_{frame} \setminus ATs_{\neg frame})$

(G) $FTs_{frame} \cup ATs_{frame}$

(H) $(FTs_{frame} \setminus FTs_{\neg frame}) \cup (ATs_{frame} \setminus ATs_{\neg frame})$

For each approach, we optionally remove the topic and the individual words the topic is constituted of from the set of candidates (i. e., $\setminus topic$) and remove the terms *FTs* extracted over the complete collection of arguments (i. e., $\setminus FTs_{corpus}$). Removing the topic and the most frequent terms within the topic, follows the idea of topic-removal for aspect-based argument clustering by Ajjour et al. (2019). Moreover, this way, we can make sure that terms belonging to the topic are not utilized for describing a frame. Afterwards, we weight the candidates per cluster applying class-based TF-IDF (c-TF-IDF) (Grootendorst, 2022). The next step is optional and filters the set of candidates by applying maximal marginal relevance (MMR) (Carbonell and Goldstein, 1998) as in the topic model BERTopic (Grootendorst, 2022) and refines the weighting of the resulting candidates by applying Jensen-Shannon divergence (JSD) as in Carmel et al. (2009). Finally, the name for a frame is either the list of top-i candidates based on the raking by their weight or the name generated by prompting a large language model (LLM) with the task of naming the frame / "subtopic"[4] given the set of all candidates sorted by their weight.

5 Evaluation

In the following, we individually evaluate the identification and the naming of frames. For each of these subtasks, we detail our experimental setup and subsequently present the results and a discussion of these.

5.1 Dataset and Preprocessing

The Argument Aspect Corpus (AAC)[5] (Ruckdeschel and Wiedemann, 2022) is built for the task of frame detection on the token-level and can thus be viewed as a ground truth dataset for our task. To our knowledge, it is the only dataset containing argumentative sentences and following our definition of overlapping argumentation frames including their naming. The AAC is based on the argumentative sentences written in English of the UKP Sentential Argument Mining Corpus (Stab et al., 2018) regarding the topics abortion, marijuana legalization, minimum wage, and nuclear energy. These topics are not part of the UKP ASPECT Corpus (Reimers et al., 2019) that was used for fine-tuning the embeddings. Per topic, there are $1,118$ to $1,502$ arguments and 12 to 15 frames excluding the frame with the label *Other*. Note, that the definition of frames utilized in this work precludes the usage of an *Other* category, which is a grouping of multiple unrelated aspects. Hence, we include all arguments of the AAC for identifying frames but ignore the *Other* label and arguments solely belonging to it during evaluation and for naming the frames. The overlap size of the frames in the AAC dataset is 1.2261, meaning that approximately every fifth argument belongs not only to one frame but to two frames.

5.2 Identification of Frames

Experimental Procedure Based on the results in Schindler (2024), we apply the following dimensionality reduction prior to the clustering: Principal component analysis (PCA) (Pearson, 1901) with 75 components on the term-level and uniform manifold approximation and projection (UMAP) (McInnes and Healy, 2018) with 50 output dimensions and a local neighborhood of 30 on the argument-level.

We perform the topic-wise grouping of arguments into frames by applying the overlapping clustering algorithm (i. e., FCM) on the argument-level and the exclusive clustering algorithm (i. e., k-

[4] While subtopics and frames are generally distinct concepts, subtopic is a more well-known and well-defined term and due to the limited context, we assume subtopics and frames to be equivalent in the scope of this task.

[5] https://doi.org/10.5281/zenodo.7525183

means) on the term-level. The clustering on the term-level is mapped back to the argument-level based on the association of the terms to the arguments. To compare against the so-far common procedure, we additionally apply exclusive clustering with k-means to the argument-level.

Building upon the PyClustering library (Novikov, 2019), we transform the results of FCM into a hard overlapping clustering by either setting $\theta_{\text{cum}} = 0$, which equals selecting the cluster(s) with the highest score, or $\theta_{\text{cum}} = 0.5$. The initial points for the clustering are selected by the k-means++ algorithm (Arthur and Vassilvitskii, 2007) utilizing the farthest points as centers. Since the amount k of frames is not known in advance, we test different values between 6 and 21. Based on previous works (Boydstun et al., 2014; Dumani et al., 2021; Jurkschat et al., 2022; Ruckdeschel and Wiedemann, 2022; Aicher et al., 2022), where frames and generic aspect-based categories were defined manually, this is a reasonable range. We are not aware of a common method to automatically determine the amount of clusters for an overlapping clustering without accessing the ground truth. Therefore, we report our results averaged over all k to get an insight on the overall performance independent of the selection of k.

Since we evaluate the identification of frames on the argument-level, where overlapping clusters are formed, we apply measures suited for this kind of clustering. Following (N'Cir et al., 2015), we report the extrinsic measures P_{sim}, R_{sim}, $F1_{sim}$, $BCubed$-P_{sim}, $BCubed$-R_{sim}, and $BCubed$-$F1_{sim}$ and the intrinsic measure *overlap size*. The $BCubed$- variants are calculated with a re-implementation[6] of the work by (Amigó et al., 2009), extending the measures from the domain of exclusive to overlapping clustering. An advantage of the $BCubed$- measures over the regular ones is that they additionally consider the amount of predicted and ground truth clusters shared between the pairs of arguments. For consistency with formerly reported measures on the task of aspect-based argument similarity prediction (Reimers et al., 2019), we additionally calculate P_{dissim}, R_{dissim}, $F1_{dissim}$ and $F1_{marco}$. Moreover, we report *OmegaSoft*[7] (Lutov et al., 2019), which is a generalization of the adjusted Rand index (ARI) for overlapping clusters, and *GNMI*[8] (Lutov et al., 2019), which is the respective generalization of normalized mutual information (NMI). To determine the best approach for identifying frames, we focus on the measures $F1_{marco}$, $BCubed$-$F1_{sim}$, *OmegaSoft*, *GNMI*, and *overlap size*.

	term-level	argument-level		
clustering	k-means	k-means	FCM$_{(\theta_{\text{cum}}=0)}$	FCM$_{(\theta_{\text{cum}}=0.5)}$
$\mathbf{F1_{macro}}$	0.5712	0.6060	0.6075	**0.6185**
$F1_{sim}$	0.3500	0.3261	0.3284	**0.3625**
P_{sim}	0.2930	0.4843	**0.4872**	0.4378
R_{sim}	**0.4593**	0.2608	0.2627	0.3281
$F1_{dissim}$	0.7925	0.8860	**0.8865**	0.8744
P_{dissim}	**0.8566**	0.8456	0.8460	0.8531
R_{dissim}	0.7421	0.9323	**0.9330**	0.8991
$\mathbf{BCubed\text{-}F1_{sim}}$	0.3564	0.3735	0.3738	**0.4115**
$BCubed$-P_{sim}	0.2931	0.4969	**0.4981**	0.4691
$BCubed$-R_{sim}	**0.4766**	0.3109	0.3108	0.3804
OmegaSoft	0.1467	0.2255	0.2283	**0.2377**
GNMI	0.2316	0.3941	0.3951	**0.4158**
overlap size	1.8960	1.000	1.0000	**1.2133**

Table 1: Results for identifying frames averaged over five random seeds, 16 different k, and the four topics of the AAC. The highest standard errors for the individual averaging steps are 0.0361 for the seeds, 0.0244 for k, and 0.0264 for the topics. The ground truth overlap size is 1.2261.

Results and Discussion The results averaged over five random seeds, the 16 different values of k, and the four topics of the AAC are presented in Table 1. Performing the clustering on the term-level leads to worse results than following the so-far common approach of exclusively clustering on the argument-level. The higher $(BCubed$-$)R_{sim}$ and P_{dissim} on the term-level show that more arguments are regarded as similar on the term-level than on the argument-level. On the argument-level, FCM$_{(\theta_{\text{cum}}=0.5)}$ is performing slightly better than FCM$_{(\theta_{\text{cum}}=0)}$ by 0.94 up to 3.77 percent points, while FCM$_{(\theta_{\text{cum}}=0)}$ and k-means are performing equally well. Moreover, we can observe the following relationship between the performance of the approaches and their *overlap size*: The closer the *overlap size* is to the ground truth, the better the performance of the approach.

This observation can be explained by the indicative role of the overlap size for the amount of argument pairs that are regarded as similar or dissimilar. The higher the overlap in the clustering, the more arguments are predicted to be similar in terms of the aspects they address. Thus, it is not surprising that the algorithms show a better performance, the more this property is in line with the data that we compare against. A reason why the term-level is not performing as well as the argument-level, could

[6] https://github.com/hhromic/python-bcubed

[7] https://github.com/eXascaleInfolab/xmeasures

[8] https://github.com/eXascaleInfolab/GenConvNMI

be the fewer amount of context that is provided by aspect terms compared to a whole sentence. Moreover, it is possible that we identified valid frames which are differing from the ones in the dataset. Hence, our evaluation procedure comparing against the frames in the AAC, which were created with a single topic-wise pre-defined set of frame labels, might underestimate the performance. Deep clustering algorithms (Zhou et al., 2022), which are learning the embedding and the clustering of the arguments jointly, are an interesting direction for future work. While different embeddings and clustering algorithms could have been employed in this work to gain even better results, note that this was not the goal of our evaluation. Instead, we have shown that clustering arguments in an overlapping manner can overcome limitations of and improve upon the so-far common procedure of exclusively clustering arguments into frames. To this end, we utilized embeddings with human-like performance in the task of aspect-based similarity detection and a well-known centroid based clustering algorithm which is used in its exclusive formulation in related work as well.

5.3 Naming of Frames

Experimental Procedure We evaluate the automatic naming of frames by applying our set of methods topic-wise to the ground truth frames of the AAC. The aspect terms (ATs) are extracted by the summetix API[9] (Schiller et al., 2021). For the terms *FTs*, we consider 1- to 4-grams (Hoppe, 2010) with a document frequency between 0.1 and 0.9. The implementation of c-TF-IDF, MMR, and the name generation with the LLM *flan-T5-base* (Chung et al., 2022) follow the one in BERTopic (Grootendorst, 2022).

Since automatically evaluating the naming against the ground truth with exact matching is too restrictive and collecting any possible equally correct frame labels is not feasible, we conduct an annotation study. To reduce the set of approaches to a reasonable amount for the human, quantitative evaluation, we first perform a qualitative evaluation with the following criteria based on the top-1 candidate: Every frame should have a different name, otherwise the frames would be describing the same aspect category and hence could be merged. We refer to this criterion as the *diff-criterion*. Moreover, no frame must be named with (a) the name of another frame as this name is definitely wrong, (b) the topic of the arguments as this is the wrong level of granularity, or (c) with no name in case the set of candidate terms is empty. Approaches that are not fulfilling this criterion are viewed as invalid.

In the subsequent human annotation, we ask seven participants the following four questions per frame / "subtopic"[10] in the light of the broader main topic, where question (2) and (4) are rated on a seven-point Likert scale from 1 (totally disagree) to 7 (totally agree):

(1) Which of the following lists of terms describes the subtopic <*frame label*> the best?

(2) The list of terms I have selected in the previous question describes the subtopic <*frame label*> well.

(3) Which of the following lists of terms describes the subtopic <*frame label*> the worst?

(4) The list of terms I have selected in the previous question describes the subtopic <*frame label*> well.

Additionally, the participants were instructed to take the order of the list of terms into account when choosing the best and worst one. To avoid bias in the single choice questions due the order in which the lists of terms of the different approaches are presented, we randomize their sequence for every participant. We perform two plausibility checks on the annotations, more precisely per frame (a) the selected list of terms has to be different for question (1) and question (3) and (b) the rating in question (4) must not be higher than the rating in question (2). Among the plausible annotations, we select the three most agreeing ones for evaluation based on the inter-rater reliability assessed through Krippendorff's alpha (Krippendorff, 2019) for ordinal data. Therewith, we gain a result that is as objective as possible by eliminating outliers (Wachsmuth et al., 2017). For questions (1) and (3), we perform the majority vote and report the percentage of best and worst rated namings per approach. In case the majority vote is inconclusive, we do not consider any of the lists of terms as best or worst, respectively. For questions (2) and (4), we report the mean based on the averaged rating per question.

[9] https://api.summetix.com/

[10] While subtopics and frames are generally distinct concepts, subtopic is a more well-known and well-defined term and due to the limited context, we assume subtopics and frames to be equivalent in the scope of this task.

Results and Discussion Any approach building upon candidate set (F), applying JSD, or utilizing the generative approach in the last step, were not able to fulfill the diff-criterion for any topic, i. e., they were not able to produce a naming without giving at least one name to more than one frame. The approaches including JSD or the generative approach are also the only approaches producing certainly wrong names for the frames by suggesting frame labels that belong to other frames. Additionally, set (E) and (F) are prone to producing empty candidate lists. Continuing with the remaining approaches, there is no difference in our criteria for applying MMR or not. Applying both $\backslash topic$ and $\backslash FTs_{corpus}$ has the same effect as performing $\backslash topic$ or $\backslash FTs_{corpus}$ on its own, except for the topic *nuclear energy* in case of the latter. To make sure that the topic cannot be utilized as a name for a frame, we therefore propose to apply $\backslash topic$ and if the topic was not known, $\backslash FTs_{corpus}$ as an approximation of the same. With this configuration, candidate set (B) violates the diff-criterion one time and candidate set (D) three times over all topics, while candidate sets (A), (C), (G), and (H) never violate the diff-criterion.

Since it might be hard to grasp the concept of a frame by just having a look at the candidate with the highest weight, we perform the annotation study with the top-3 ranked candidates and thus include all of the six remaining sets. Therewith, our annotation study reduces to identifying the best set of candidate terms when applying $\backslash topic$ and weighting with c-TF-IDF. Our plausibility checks lead to the exclusion of two study participants. The remaining five participants have an inter-rater agreement of 0.39 for questions (1) and (3), and of 0.56 for the questions with Likert scales. The three most agreeing annotators are the same for both kinds of questions and have an agreement of 0.59 and 0.70, respectively. For the lists of terms, the descriptive fit with respect to the frame label is rated on average with 5.94 for the best and 3.33 for the worst one. This indicates that the approaches are in general able to produce a naming that is describing the frame very well, while at the same time the worst namings have a rather bad descriptiveness. Based on the evaluation of the single choice questions, which is provided in Table 2, set (B) is performing the best and set (D) is by far the worst. Thus, we can conclude that for the top-3 terms as a naming, the best configuration among our approaches is to extract the aspect terms ATs_{frame} of the frame, apply topic-removal $\backslash topic$, and weight the terms with the c-TF-IDF procedure. The naming generated with this approach and set (A) is exemplarily shown in Table 3 in the appendix.

set	A	B	C	D	G	H
best	0.44	**0.65**	0.44	0.12	0.56	0.50
worst	0.19	0.17	0.19	**0.71**	0.12	0.13

Table 2: Results for naming the frames. We report the ratio of best and worst rated list of terms over all topics for the respective approaches. Set (A) and (C) are based on frequent terms, set (B) and (D) on aspect terms, and set (G) and (H) on the union of both kind of terms.

The results indicate that the approaches with JSD or the generative naming approach are performing on the wrong level of granularity since the predicted names are more related to the general concept of the topic. For MMR, we did not observe a difference even within the top-10 since for the valid approaches, the amount of extracted candidates either is already below the 10 candidates that MMR is filtering for or MMR removed candidates that are not within the top-10 anyways. The differences in inter-coder agreement in the annotation study shows that there is still subjectivity in the ratings. However, by performing the evaluation with the three best agreeing participants, we were able to substantially improve the reliability of our results. Interestingly, the best (ATs_{frame}) as well as the worst ($ATs_{frame} \setminus ATs_{\neg frame}$) performing set are based on aspect terms. This observation allows to conclude that the aspect terms shared among different frames are highly relevant to the success of naming the frames. While this might be surprising in the first place, the context provided by the other aspect terms of the frame can lead to a different interpretation of the same aspect terms and therefore give rise to a different frame label. In future work, the approaches for naming the frames could benefit from utilizing external sources that are not generative as in this work to group the candidate terms into the underlying concept they are describing.

6 Conclusion and Future Work

We introduced a sequential three-step pipeline that not only identifies but also names frames while considering the fact that an argument can belong to more than one frame. Through evaluating each step of the pipeline individually, we obtain the following configuration: The pipeline operates on the argument-level, where the arguments are embedded together with their topic by an SBERT model that is fine-tuned for aspect-based similarity. Afterwards, we apply fuzzy c-means clustering and perform a transformation to a hard overlapping clustering such that the *overlap size* of the resulting frames is close to 1.2. Our alternative approach performing the clustering on the term-level with k-means and mapping the results back to the argument-level is performing worse than the so-far common procedure of exclusive clustering on the argument-level. The naming of the frames, which is the last step in the pipeline, is performed for each frame individually. Per frame, we select the candidates with the highest c-TF-IDF scores from a set of candidates obtained through aspect term extraction and removing any terms that are part of the topic.

The next step with respect to evaluation is to investigate the pipeline in its entirety since there is an interaction between identifying and naming frames. The data-driven identification of frames, as performed in the herein work, poses the risk of resulting in an infeasibly large amount of clusters or clusters not representing meaningful and well-defined frames (Jurkschat et al., 2022; Ruckdeschel and Wiedemann, 2022). Nevertheless, the latter is a general risk when not defining the frame labels by hand and the amount of clusters can be defined by setting the hyperparameters respectively. Though, the question remains how many frames to create. To this end, we propose to either investigate internal clustering measures or to perform a selection with a human-in-the-loop setting based on the predicted frame labels of the frames. Moreover, it would be interesting to investigate deductive approaches to frame identification as a post-processing step once the respective names of the frames are known.

Limitations

In our experiments, we did not select the amount of frames and instead averaged over all tested k. For exclusive clustering algorithms, k can be selected by applying the elbow method, average silhouette approach, Hartigan statistics, or gap static, for example (Yang et al., 2019). Though, these selection methods are still having weaknesses that need to be overcome, making the problem relevant to active research (Yang et al., 2019). Overlapping clustering algorithms, additionally, are lacking intrinsic evaluation metrics that go beyond the measure of *overlap size*. Due to this, it is not clear how to perform the selection of k without having access to the ground truth. Since an investigation of strategies for automatically determining the amount of clusters in overlapping clustering goes beyond the scope of this work, we performed an evaluation that remains agnostic to the ground truth in the dataset by treating k as a factor to be averaged out. Moreover, not selecting k based on intrinsic cluster evaluation metrics also provides the chance to have the user decide on the amount and hence the granularity of the frames that is best suited to their application. Such a decision could be guided by the respective naming of the frames for different k.

To evaluate the automatic naming of frames, we performed a qualitative evaluation on the identified ground-truth frames. Though, it would be interesting to apply the naming to the frames identified by our approach and therewith go a step towards evaluating the pipeline end-to-end. Additionally, instead of selecting the three top-ranked terms from the weighted list of candidates as the herein work, one could also investigate other heuristics for choosing the terms, assuring that the frame is well described and potentially can be transformed into a single-term description for the frame label.

Throughout the work, we based our experiments on a single dataset since to the best of our knowledge this is the only one suiting our task with topic-specific overlapping argumentation frames including ground truth labels for their naming. However, since the dataset is covering four different topics, our results should be generalizable with respect to the topic of the arguments.

Ethical Considerations

Consideration of ethics is essential for applications that work with arguments. Our proposed automation of identifying and naming frames, is a generic approach that allows to further structure collections of topic-related arguments based on the aspects they address. Employed in combination with argument search, for example, this will make discussions and their arguments more intuitively accessible to humans. Moreover, by creating awareness of frames, our work allows to, among others, discover biases and filter bubbles (Pariser, 2011; Ekström et al., 2022) in the data and thus paves the path for approaches to mitigate these. While our work could be misused to influence people, e. g., by reinforcing such biases and filter bubbles, we see the positives of our work prevailing, namely being a tool providing transparency about the frames that are existing in the data. As such, it could also be utilized in the process of discovering such malicious intentions.

Acknowledgments

This work has been funded by the DFG within the project “BEA - Building Engaging Argumentation”, Grant no. 313723125, as part of the Priority Program “Robust Argumentation Machines (RATIO)” (SPP-1999). We thank summetix GmbH for supporting our research with access to their Aspect API.

References

Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. 2001. On the surprising behavior of distance metrics in high dimensional spaces. In *Database Theory - ICDT 2001, 8th International Conference, London, UK, January 4-6, 2001, Proceedings*, volume 1973 of *Lecture Notes in Computer Science*, pages 420–434. Springer.

Annalena Aicher, Nadine Gerstenlauer, Isabel Feustel, Wolfgang Minker, and Stefan Ultes. 2022. Towards building a spoken dialogue system for argument exploration. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1234–1241, Marseille, France. European Language Resources Association.

Annalena Aicher, Niklas Rach, Wolfgang Minker, and Stefan Ultes. 2019. Opinion building based on the argumentative dialogue system BEA. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction - 10th International Workshop on Spoken Dialogue Systems, IWSDS 2019, Syracuse, Sicily, Italy, 24-26 April 2019*, volume 714 of *Lecture Notes in Electrical Engineering*, pages 307–318. Springer.

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling frames in argumentation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2922–2932, Hong Kong, China. Association for Computational Linguistics.

Yamen Ajjour, Henning Wachsmuth, Dora Kiesel, Patrick Riehmann, Fan Fan, Giuliano Castiglia, Rosemary Adejoh, Bernd Fröhlich, and Benno Stein. 2018. Visualization of the topic space of argument search results in args.me. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 60–65, Brussels, Belgium. Association for Computational Linguistics.

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr.*, 12(4):461–486.

David Arthur and Sergei Vassilvitskii. 2007. k-means++: the advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007*, pages 1027–1035. SIAM.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020a. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4029–4039, Online. Association for Computational Linguistics.

Roy Bar-Haim, Yoav Kantor, Lilach Eden, Roni Friedman, Dan Lahav, and Noam Slonim. 2020b. Quantitative argument summarization and beyond: Cross-domain key point analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 39–49, Online. Association for Computational Linguistics.

Roy Bar-Haim, Yoav Kantor, Elad Venezian, Yoav Katz, and Noam Slonim. 2021. Project Debater APIs: Decomposing the AI grand challenge. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 267–274, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

James C. Bezdek, Robert Ehrlich, and William Full. 1984. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2-3):191–203.

Amber E. Boydstun, Dallas Card, Justin Gross, Paul Resnick, and Noah A. Smith. 2014. Tracking the Development of Media Frames within and across Policy Issues.

Jaime G. Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 335–336. ACM.

David Carmel, Haggai Roitman, and Naama Zwerdling. 2009. Enhancing cluster labeling using wikipedia. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 139–146, New York, NY, USA. Association for Computing Machinery.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, and 12 others. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Rob Churchill and Lisa Singh. 2022. The evolution of topic modeling. *ACM Comput. Surv.*, 54(10s):215:1–215:35.

Johannes Daxenberger, Benjamin Schiller, Chris Stahlhut, Erik Kaiser, and Iryna Gurevych. 2020. Argumentext: Argument classification and clustering in a generalized search scenario. *Datenbank-Spektrum*, 20(2):115–121.

Lorik Dumani, Tobias Wiesenfeldt, and Ralf Schenkel. 2021. Fine and coarse granular argument classification before clustering. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 422–432. ACM.

Axel G. Ekström, Diederick C. Niehorster, and Erik J. Olsson. 2022. Self-imposed filter bubbles: Selective attention and exposure in online search. *Computers in Human Behavior Reports*, 7:100226.

Robert M. Entman. 1993. Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication*, 43(4):51–58.

Michael Färber and Anna Steyer. 2021. Towards full-fledged argument search: A framework for extracting and clustering arguments from unstructured text. *CoRR*, abs/2112.00160.

Roni Friedman, Lena Dankin, Yufang Hou, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2021. Overview of the 2021 key point analysis shared task. In *Proceedings of the 8th Workshop on Argument Mining*, pages 154–164, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based TF-IDF procedure. *CoRR*, abs/2203.05794.

Shohreh Haddadan, Elena Cabrio, Axel J. Soto, and Serena Villata. 2022. Topic modelling and frame identification for political arguments. In *AIxIA 2022 - Advances in Artificial Intelligence - XXIst International Conference of the Italian Association for Artificial Intelligence, AIxIA 2022, Udine, Italy, November 28 - December 2, 2022, Proceedings*, volume 13796 of *Lecture Notes in Computer Science*, pages 268–281. Springer.

Philipp Heinisch and Philipp Cimiano. 2021. A multi-task approach to argument frame classification at variable granularity levels. *it Inf. Technol.*, 63(1):59–72.

Dennis Hoppe. 2010. Cluster-labeling: Paradigmen und validierung. Master's thesis, Bauhaus-Universität Weimar, Fakultät Medien, Medieninformatik.

Lena Jurkschat, Gregor Wiedemann, Maximilian Heinrich, Mattes Ruckdeschel, and Sunna Torge. 2022. Few-shot learning for argument aspects of the nuclear energy debate. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 663–672, Marseille, France. European Language Resources Association.

Klaus Krippendorff. 2019. *Content Analysis: An Introduction to Its Methodology*, 4 edition. SAGE Publications, Inc.

Sha Lai, Yanru Jiang, Lei Guo, Margrit Betke, Prakash Ishwar, and Derry Tanti Wijaya. 2022. An unsupervised approach to discover media frames. In *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*, pages 22–31, Marseille, France. European Language Resources Association.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159–165.

Artem Lutov, Mourad Khayati, and Philippe Cudré-Mauroux. 2019. Accuracy evaluation of overlapping and multi-resolution clustering algorithms on large datasets. In *IEEE International Conference on Big Data and Smart Computing, BigComp 2019, Kyoto, Japan, February 27 - March 2, 2019*, pages 1–8. IEEE.

James MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press.

Leland McInnes and John Healy. 2018. UMAP: uniform manifold approximation and projection for dimension reduction. *CoRR*, abs/1802.03426.

Amita Misra, Brian Ecker, and Marilyn Walker. 2016. Measuring the similarity of sentential arguments in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 276–287, Los Angeles. Association for Computational Linguistics.

Xinyi Mou, Zhongyu Wei, Changjian Jiang, and Jiajie Peng. 2022. A two stage adaptation framework for frame detection via prompt learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2968–2978, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Chiheb-Eddine Ben N'Cir, Guillaume Cleuziou, and Nadia Essoussi. 2015. *Overview of Overlapping Partitional Clustering Methods*, pages 245–275. Springer International Publishing, Cham.

Andrei Novikov. 2019. PyClustering: Data mining library. *Journal of Open Source Software*, 4(36):1230.

Juri Opitz, Philipp Heinisch, Philipp Wiesenbach, Philipp Cimiano, and Anette Frank. 2021. Explainable unsupervised argument similarity rating with Abstract Meaning Representation and conclusion generation. In *Proceedings of the 8th Workshop on Argument Mining*, pages 24–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

E. Pariser. 2011. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin Publishing Group.

Karl Pearson. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Niklas Rach, Klaus Weber, Louisa Pragst, Elisabeth André, Wolfgang Minker, and Stefan Ultes. 2018. EVA: A multimodal argumentative dialogue system. In *Proceedings of the 2018 on International Conference on Multimodal Interaction, ICMI 2018, Boulder, CO, USA, October 16-20, 2018*, pages 551–552. ACM.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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*, pages 567–578, Florence, Italy. Association for Computational Linguistics.

Mattes Ruckdeschel and Gregor Wiedemann. 2022. Boundary detection and categorization of argument aspects via supervised learning. In *Proceedings of the 9th Workshop on Argument Mining*, pages 126–136, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396, Online. Association for Computational Linguistics.

Carolin Schindler. 2024. Overlapping aspect-based argument cluster analysis including cluster labelling for opinion formation with argumentative dialogue systems. Master's thesis, Institute of Communications Engineering, University of Ulm.

Bernd Skiera, Shunyao Yan, Johannes Daxenberger, Marcus Dombois, and Iryna Gurevych. 2022. Using information-seeking argument mining to improve service. *Journal of Service Research*, 25(4):537–548.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, and 34 others. 2021. An autonomous debating system. *Nat.*, 591(7850):379–384.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Michael Steinbach, Levent Ertöz, and Vipin Kumar. 2004. *The Challenges of Clustering High Dimensional Data*, pages 273–309. Springer Berlin Heidelberg, Berlin, Heidelberg.

Shahbaz Syed, Timon Ziegenbein, Philipp Heinisch, Henning Wachsmuth, and Martin Potthast. 2023. Frame-oriented summarization of argumentative discussions. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 114–129, Prague, Czechia. Association for Computational Linguistics.

Dietrich Trautmann. 2020. Aspect-based argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 41–52, Online. Association for Computational Linguistics.

Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2020. Fine-grained argument unit recognition and classification. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9048–9056. AAAI Press.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association* *for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Jaekyung Yang, Jong-Yeong Lee, Myoungjin Choi, and Yeongin Joo. 2019. A new approach to determine the optimal number of clusters based on the gap statistic. In *Machine Learning for Networking - Second IFIP TC 6 International Conference, MLN 2019, Paris, France, December 3-5, 2019, Revised Selected Papers*, volume 12081 of *Lecture Notes in Computer Science*, pages 227–239. Springer.

Sheng Zhou, Hongjia Xu, Zhuonan Zheng, Jiawei Chen, Zhao Li, Jiajun Bu, Jia Wu, Xin Wang, Wenwu Zhu, and Martin Ester. 2022. A comprehensive survey on deep clustering: Taxonomy, challenges, and future directions. *CoRR*, abs/2206.07579.

A Example for Naming of Frames

Table 3 shows the lists of top-3 candidate terms for set (A) and (B) as utilized in the annotation study for the topics *abortion* and *minimum wage*. The configuration with set (B) was identified as the best performing one for naming the frames.

ground truth	set (A): FTs_{frame}	set (B): ATs_{frame}
abortion		
abortion industry	industry, profit, consistent	industry, profit, dirty
adoption	adoption, couple, adopt	adoption, baby, kid
bodily autonomy/women's rights	choice, body, make	choice, body, decision
consequences of childbirth	welfare, unwanted, care	raise, unwanted, poverty
fetal defects/disabilities	defect, syndrome, fetal	syndrome, health, pregnancy
fetal/newborn rights	fetus, person, unborn	fetus, person, unborn
funding of abortion	poor, medicaid, funding	poor, medicaid, funding
health effects of pregnancy/childbirth	pregnancy, mother, risk	pregnancy, mother, risk
illegal abortions	illegal, unsafe, 000	illegal, unsafe, alley
moral/ethical values	god, moral, immoral	moral, religious, catholic
parental consent	parental, minor, parent	minor, parent, consent
psychological effects of abortion	regret, mental, psychological	regret, mental, psychological
rape	rape, incest, raped	rape, incest, raped
responsibility	contraception, control, use	control, contraception, contraceptive
safety/health effects of legal abortion	cancer, breast, risk	cancer, risk, medical
minimum wage		
capital vs labor	power, sweatshop, bargaining	market, labor, monopsony
competition/business challenges	small, company, owner	small, profit, hotel
economic impact	economy, spend, money	economy, spend, money
government	government, market, free	government, market, state
low-skilled	skilled, unskilled, employment	employment, young, skill
motivation/chances	school, opportunity, skill	school, opportunity, skill
prices	price, consumer, raise	price, consumer, raise
social justice/injustice	poverty, living, income	poverty, income, inflation
turnover	turnover, training, employee	turnover, productivity, reduce
un/employment rate	employment, unemployment, effect	employment, unemployment, labor
welfare	tax, program, assistance	tax, government, income
youth and secondary wage earners	household, family, teenager	household, family, teenager

Table 3: Frame labels as predicted by the automated naming approaches selecting the top-3 candidates ranked by their c-TF-IDF weight from the respective candidate set with topic-removal.