Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 663–672
Marseille, 20-25 June 2022

Few-Shot Learning for Argument Aspects of the Nuclear Energy Debate

Lena Jurkschat[1], Gregor Wiedemann[2], Maximilian Heinrich[3], Mattes Ruckdeschel[2], Sunna Torge[1]
[1]Technische Universität Dresden, Germany
[2]Leibniz-Institute for Media Research | Hans-Bredow-Institute, Germany
[3]Leipzig University, Germany
{lena.jurkschat, sunna.torge}@tu-dresden.de, mheinrich@informatik.uni-leipzig.de
{g.wiedemann, m.ruckdeschel}@leibniz-hbi.de

Abstract

We approach aspect-based argument mining as a supervised machine learning task to classify arguments into semantically coherent groups referring to the same defined aspect categories. As an exemplary use case, we introduce the *Argument Aspect Corpus–Nuclear Energy* that separates arguments about the topic of nuclear energy into nine major aspects. Since the collection of training data for further aspects and topics is costly, we investigate the potential for current transformer-based few-shot learning approaches to accurately classify argument aspects. The best approach is applied to a British newspaper corpus covering the debate on nuclear energy over the past 21 years. Our evaluation shows that a stable prediction of shares of argument aspects in this debate is feasible with 50 to 100 training samples per aspect. Moreover, we see signals for a clear shift in the public discourse in favor of nuclear energy in recent years. This revelation of changing patterns of pro and contra arguments related to certain aspects over time demonstrates the potential of supervised argument aspect detection for tracking issue-specific media discourses.

Keywords: argument mining, aspect-based argument mining, argument frames, argument aspects, text classification, few-shot learning, nuclear energy discourse

1. Mining for Argument Aspects

The field of argument mining has gained increasing attention in natural language processing during the past decade (Lawrence and Reed, 2020). The automatic identification and extraction of argumentative structures promises not only commercial opportunities but also scientific innovation in applying disciplines. For instance, the continuous monitoring of public debates in social and mass media regarding the distribution and development of argumentation about certain societal issues such as nuclear energy is of enormous interest, especially in the field of social and communication sciences. The nuclear energy debate, for instance, recently reached a new climax along with the question of whether this technology can be considered sustainable energy in the face of climate change threats. However, unlike sentiment analysis, which has been widely adopted in social research, argument mining still is rarely used. Despite the impressive progress and the increasing availability of language resources as training data, we assume two main reasons for the low prevalence of argument mining applications.

First, from the perspective of computational linguistics argument mining has been treated to a large extent as a structuring problem of unstructured language. Most works focus on sub-tasks such as the automatic identification of argumentative units in natural language texts, their separation into components such as premises or claims, and their relation of either supporting or attacking each other. Nevertheless, for public discourses in social networks and mass media, it became evident that argumentative utterances only rarely follow a formal argumentative structure (e.g. premise and claim) or refer to each other explicitly. Instead, argumentation in public debates is largely governed by implicit assumptions, common sense knowledge, isolated expressions of stances, and claims without explicit justification or discourse markers (Moens, 2018). This circumstance sparked a growing interest in the detection of specific claims (Lapesa et al., 2020) as well as the formal modeling of argument semantics (Baumann et al., 2020).

Second, the generalizability and domain adaptation of already published language resources is either not sufficient, or sufficiently tested for the application in media studies. Moreover, so far the field lacks manually annotated corpora of issue-specific semantic argument components. The analysis of media discourse requires not only the identification of structural components and stances on particular issues but also finer-grained information about aspects that repeatedly address the same problem or argument. So far, this problem has been approached only in an unsupervised manner by data-driven clustering of arguments (Heinisch and Cimiano, 2021; Ajjour et al., 2019). These approaches, however, often end up with an unmanageably large number of clusters, many of which do not represent actual argument aspects. In addition, they conflict with the deductive research paradigm common in social science and media research that requires well-defined, theoretically or empirically grounded categories.

In contrast, we approach *aspect-based argument mining* (ABAM) as a supervised sub-task of argument mining to obtain issue-specific aspect categories for arguments. There are three main contributions of our paper:

1. We introduce the *Argument Aspect Corpus–Nuclear Energy* (AAC-NE) containing English-

language sentences with aspect annotations describing the content of arguments from the topic of nuclear energy.

2. Since this supervised approach requires issue-specific aspect category sets and training data, we further apply recently published few-shot classification approaches to evaluate opportunities to train aspect classifiers more cost-efficiently with less training data.

3. In a final case study, we apply the best classifier on a diachronic newspaper dataset to test the stability of aspect predictions in a few-shot setting. This study also reveals some interesting changes of argument patterns over time.

In Section 2, we present the aforementioned related work in more detail and briefly introduce two methods of few-shot learning that will be employed in our study. Then, Section 3 describes our process of developing an aspect category set based on similarity-based clustering of an issue-labeled argument dataset, and group discussions. In Section 4, we perform experiments to evaluate different transformer neural networks and the few-shot learning approaches building up on them. In Section 5, we apply the best approach to a corpus of British newspaper articles covering the last 21 years to evaluate the stability of few-shot learning-based predictions and to demonstrate the application potential of supervised ABAM, followed by a brief summary of our work in Section 6.

2. Related Work

The work presented in this paper builds upon two main research fields: *argument mining* with a special focus on argument aspects, and *few-shot learning* for text classification.

Throughout the recent literature, different conceptualizations of **argument aspects** have been proposed based on argument similarity, aspect terms, and distant supervision. Also, a few corpora are available for the English language. First, Misra et al. (2016) introduced the *Argument Facet Similarity Corpus* (AFS) in which pairs of argumentative units are annotated how much they address the same argument. This idea was adopted by Reimers et al. (2019), who published the UKP ASPECT dataset. In these datasets, argument aspects remain an implicit concept expressed by the crowd-sourced similarity score. While the similarity scores are useful to evaluate a data-driven clustering of arguments, they do not directly translate to well-defined aspect categories that can describe the evolution of a public debate. A similar approach to detect argument subgroups and aspects is described in Daxenberger et al. (2020). They use agglomerative hierarchical clustering on contextualized embeddings of argument pairs based on BERT (Devlin et al., 2019) and ELMO (Peters et al., 2018) embeddings. The obtained clusters, however, in many cases do not provide well-defined aspect categories, too.

To approach the actual content of arguments more directly, Trautmann (2020) proposed the task of *aspect term extraction* (ATE). Analog to aspect-based sentiment analysis, he employs a sequence tagging approach to extract tokens from argumentative texts that represent the main points to which an argument refers. In this task, no generalization of extracted main points into general aspect categories is performed. Partially, this is approached by Ajjour et al. (2019) who strive to cluster arguments that refer to the same aspect into frames by removing topic-specific features from the texts. The theoretical concept of "framing" is actually adapted from media studies. However, in their operationalization for cluster evaluation, they rely on distant supervision by grouping arguments from debate portals posted under (nearly) identical sub headings. These user-generated sub headings hardly qualify as well-defined aspects or frame categories as required by the social sciences. Heinisch and Cimiano (2021) present another approach to classify arguments into frames by fitting argument clusters based on compressed word embeddings to a set of pre-defined media frame categories (Boydstun et al., 2014). Apart from these generic media frames, however, no issue-specific aspect or frame categories are developed.

Unlike these previous works that approach ABAM either as an unsupervised clustering task or as a generic frame classification task, we believe that in order to track issue-specific media discourses, issue-specific annotated datasets in combination with supervised classifiers are required. It is not sufficient only to detect groups of arguments. In addition, there is a need for empirically grounded, well-defined aspect categories, which are yielded by data-driven approaches in common with dedicated methodologies. Therefore, in contrast to the above-mentioned work, we use unsupervised methods only in the first step to detect argument groups, based on which well-defined aspect categories are developed later on.

Few-shot learning approaches for classification recently gained a lot of attention and address the lack of training data in several ways. Current models are often built on top of the transformer-based neural network architecture and pretrained language models based thereon as introduced by Devlin et al. (2019). In Schick and Schütze (2021), the *Pattern-Exploiting Training* (PET) is introduced, which is a semi-supervised training procedure that reformulates input examples as cloze-style phrases to help language models to better understand a given task. In Halder et al. (2020), the *Task-Aware Representation of Sentences* (TARS) is presented. Here, the input consists of two separated parts: the text to be classified, and a semantically descriptive category label. The classification task itself reduces to the binary decision of whether text and label

match. In our work, we evaluate PET ensemble models and pretrained TARS models for argument aspect classification.

3. The AAC-NE Dataset

For categorizing nuclear energy arguments by their aspect, we selected a previously published language resource that already identified topic-specific argumentative sentences from a variety of web sources as a starting point. The English-language corpus UKP SAM (Stab et al., 2018) consists of 25,492 sentences assigned to eight different topics. Each sentence is either labeled as 'supporting', 'opposing' or 'non-argument' w.r.t. its specific topic. The topic of nuclear energy in this corpus consists of 740 pro and 564 contra arguments. Sentences with the 'no argument' label were not considered.

Given an argument from a specific topic, we define an aspect as a repeatedly occurring problem or related subtopic within this topic that can be distinguished from other aspects reliably. Thus, arguments addressing the same aspect overlap in terms of their content of meaning. However, the definition and granularity of aspects may vary greatly between different readers of the same set of arguments. Therefore, we decided to perform an automated pre-clustering of arguments in order to reduce subjective bias and carrying out a more systematic approach to determine aspect labels.

3.1. Clustering

To identify a suitable pre-clustering approach, we used the UKP ASPECT dataset (Reimers et al., 2019). This corpus contains 3595 sentence pairs for 28 different topics. The similarity of each sentence pair is annotated with one of the following categories 'Different Topic/Can't decide', 'No Similarity', 'Some Similarity', and 'High Similarity'.

For finding an appropriate clustering method, we tested a multi-view clustering based on Fraj et al. (2019) as well as a hierarchical single-view clustering with sentence embeddings generated by the S-BERT transformer model (Reimers and Gurevych, 2019). We compared the results with (Reimers et al., 2019), who are using a single-view clustering with BERT embeddings on the UKP ASPECT corpus. The corpus annotations were used in order to determine the quality of the clustering. It showed that the hierarchical single-view clustering on S-BERT embeddings provided the best results. We applied this procedure on all argumentative sentences of the nuclear energy topic from the UKP SAM corpus with a fixed number of clusters. We decided for a grouping into 15 clusters—a large enough number to represent various debate aspects but small enough to allow for a close, qualitative inspection.

3.2. Definition Aspects

The obtained clusters were analyzed and discussed in our research group in order to obtain the final aspects.

Aspect	Train	Val	Test	$\sum$	α
alternatives	100	16	21	137	0.69
costs	98	17	29	144	0.72
environment	209	27	64	300	0.74
innovation	33	2	8	43	0.38
reactor safety	112	17	43	172	0.59
reliability	47	5	10	62	0.36
waste	87	5	26	118	0.80
weapons	52	11	15	78	0.77
other	120	23	29	172	0.49
all	858	123	245	1226	0.62

Table 1: AAC-NE corpus statistics. The last column shows the inter-annotator agreement for each category (Krippendorff's α)

The goal was to define the aspects as precisely as possible with as little as possible overlap between them. First, we defined aspects as subtopics in the field of nuclear energy. Secondly, we tried to restrict the room for interpretations by disallowing background assumptions to label a cluster with an aspect. That is, for an aspect to be present, it must be possible to link it directly to words or sequences of words that express it, instead of just being (logically) implied by the content that is actually present. With those restrictions, we extracted 9 aspect categories out of the 15 clusters. Not every cluster represented an aspect. For instance, we found clusters that just represented stances on the subject matter or contained arguments for and against nuclear energy only in Australia. The final aspect categories are described in Table A1. Despite our efforts to create aspects systematically via a data-driven clustering, aspect definitions remain open to subjective interpretation to some extent. Additionally, we probably experienced some form of confirmation bias during the process of inferring aspects from clusters, that could not eliminated completely. Arguments might have been blanked out unconsciously if they did not fit the aspect assumed beforehand. At the end of our group discussion, we concluded that we had fully mapped all aspects that predominantly occurred in the clusters. However, another group of researchers repeating the same procedure might come up with an overlapping but slightly different set of aspect categories.

3.3. Annotation

In sentiment analysis, aspect target detection is usually operationalized as a sequence tagging task targeting one or a few tokens (e.g. the 'display quality' of a cell phone). Argument aspects, in contrast, often need longer sequences or entire sentences to be identified.[1]

[1]The ATE task for ABAM as proposed by Trautmann (2020) actually very much resembles aspect target detection as it also extracts short token sequences via sequence labeling. However, the extracted aspect terms are neither sufficient

Thus, we decided to approach supervised ABAM as a short text classification task that can be applied after argument unit segmentation in an argument mining pipeline. For the annotation process, we recruited student assistants and faculty members with an education in social sciences. Four annotators spent about 30 min for training to reduce interpretation bias. Annotations were obtained with the doccano tool (Nakayama et al., 2018) from three annotators per sentence. The final label was decided by majority vote. The inter-coder reliability for the annotation was evaluated with Krippendorff's alpha (see Table 1). Aspects that are regularly expressed with a limited vocabulary (e.g. *costs*) were annotated quite reliably. Other aspects that are expressed in significantly more diverse ways (e.g. technological *innovation*), achieve much lower inter-coder reliability.

4. Few-shot Learning for Argument Aspects

We perform two experiments on the AAC-NE dataset to identify the best performing classifier based on fine-tuning current pretrained language models. Furthermore, we evaluate three different few-shot learning approaches based on that model to learn about how aspect detection for new issues can be automated efficiently.

4.1. Fine-tuning Transformers

Parameter	Value
maximum sequence length	256
batch size	4
lr scheduler	linear
warmup ratio	0.1
learning rate	5e-6
maximum number of epochs	20

Table 2: Hyperparameters for transformer fine-tuning

Since Devlin et al. (2019) published the BERT model, fine-tuning of pretrained language models based on variants of the transformer architecture sets the state of the art for text classification. We compare six different models: ALBERT (Lan et al., 2020), BERT (Devlin et al., 2019) both in their cased base and large versions, and ELECTRA (Clark et al., 2020), RoBERTa (Liu et al., 2019), and XLM-RoBERTa (Conneau et al., 2020) in their large version. All models are fine-tuned on the AAC-NE training set with the same reasonable default hyperparameters (cp. Table 2). The best-performing model on the validation set is used for evaluation on the hold-out test set.

Table 3 reports the mean test set performance and standard deviation of each model for five repeated runs. The results indicate that the BERT base model performs surprisingly well compared to its successors with much more parameters. With 78.4 % accuracy and 74.1 % F1-score, the RoBERTa model, however, clearly outperforms the other pretrained models and will be the basis for our few-shot experiments. Table A2 reports the detailed results for individual aspect categories.

4.2. Transformer-based Few-shot Learning

In the second experiment, we evaluate how different few-shot learners are able to solve the aspect classification task with small training datasets. Training sets are increased in steps of $n \in \{1, 2, 4, 8, 10, 50, 100, all\}$ samples for each label.[2] For each training data size, the training was repeated five times to account for random effects from model initialization.

Baseline FT: To compare few-shot learners with the conventional procedure of fine-tuning, we train the RoBERTa model with training sets of each size analog to the previous experiment.

S-BERT k-NN: We calculate sentence embeddings with the model `all-mpnet-base-v2` of Sentence-BERT (Reimers and Gurevych, 2019) for each argument in the training data. We further construct a *label sentence* for each label in the form "In the nuclear energy debate, the argument is about $[label]$". For each embedding of a test sentence, we then calculated the k ($k = 3$ if $n < 4$ else $k = 5$) most similar embedding vectors in this extended training set using cosine similarity. The label that occurred most frequently among the top k results is chosen as the predicted label. In the case of a tie between several labels, the label with the highest accumulated cosine similarity wins.

TARS: Halder et al. (2020) introduce a *task-aware representation of sentences for multi-label text classification* (TARS) where a label name and a sentence are separated by a [SEP] token and fed into a transformer network. As a target, the model learns the matching between sentences and labels as a binary output. This way, the TARS model can make use of the semantic information contained in the labels of a category set which allows its application to zero- and few-shot learning scenarios. To prepare a language model for this reformulation of the multi-label classification task, it must be pretrained with data sets that are as similar as possible to the actual target task. Since, to our knowledge, there is no dataset of argument aspects, we pretrain the RoBERTa-large model in the TARS-approach with the topic information from the UKP-SAM dataset. We pretrain this TARS model with a batch size of 4 and a learning rate of 0.01 for 20 epochs. The resulting

to describe argument aspects in the sense of the repeated address of the same problem, nor can they be easily grouped into sets of terms that address the same aspect.

[2]For larger n there are not enough examples in the full training set for some labels (cp. Tab. 1). For these labels, the few-shot samples at steps 50 and 100 are identical with the full set of training examples that specific label.

	Precision		Recall		F1		Accuracy	
Model	mean	std	mean	std	mean	std	mean	std
albert-base-v2	68.5	±8.6	67.8	±5.2	67.2	±5.7	75.1	±2.0
bert-base-cased	72.2	±3.0	75.3	±3.2	71.6	±3.3	77.6	±3.4
bert-large-cased	67.1	±2.7	70.7	±4.2	67.0	±3.1	75.3	±1.7
electra-large	72.0	±2.6	73.8	±2.9	70.0	±1.7	73.8	±2.7
roberta-large	**75.6**	±4.7	**77.2**	±2.2	**74.1**	±2.5	**78.4**	±2.9
xlm-roberta-large	69.3	±3.7	69.5	±5.4	67.5	±3.6	73.5	±1.5

Table 3: Macro average performance (in %) of argument aspect classification with different transformer models.

model is the basis for few-shot learning on the AAC-NE corpus using the same hyperparameters.

PET: Schick and Schütze (2021) propose *Pattern-Exploiting Training* to perform few-shot learning that utilizes label information in form of cloze-style phrases to guide language models for classification tasks. We construct three pattern-verbalizer pairs (PVP):

- "This argument addresses the aspect $[label]$ in the nuclear energy debate. [SEP] $[text]$"
- "In the nuclear energy debate, '$[text]$' is about $[label]$."
- "This atomic energy argument refers to $[label]$. [SEP] $[text]$ "

During training, a separate model is fine-tuned for all PVPs where $[text]$ is replaced with the argument text. For the masked $[label]$ token a cross-entropy (CE) loss is calculated on the normalized probability of the predicted token and the one-hot truth vector on the label set. For this, long label names had to be replaced with single token equivalents from the RoBERTa tokenizer. Further, PET makes use of unlabeled data to combine the CE loss with a masked language model loss as a second learning target to prevent catastrophic forgetting. For this, we use 1000 argumentative sentences from the British newspaper corpus (cp. Section 5). For the final label decision, PET creates an ensemble from all separate PVP models. Due to this, PET is computationally much more expensive than TARS. We train PET with batch-size 4 and a learning rate of 1e-5 for 10 epochs.

Table 4 shows the classification performance for the baseline of fine-tuning RoBERTa and the three different few-shot learning approaches at different training set sizes. Basic fine-tuning, as expected, fails on small datasets. In contrast, the k-NN approach based on S-BERT embeddings works surprisingly well for very small training data but does not benefit from larger training sets compared to the other approaches. With no actual training or fine-tuning, the approach seems to be very dependent on the textual similarity between training and test samples. This works especially well for aspects such as *costs* or *weapons* that are usually verbalized with very narrow vocabulary. The TARS model outperforms the fine-tuning baseline also only for very small training sets. In the original paper, the approach heavily depends on pretraining on tasks with similar label sets. Due to the absence of such a dataset for our ABAM task, TARS does not seem to be very useful in our scenario. The best performance and most stable results for medium training sets are achieved by the PET model with macro-F1 scores > 67 % for 8 and more training examples.

	Baseline FT			S-BERT k-NN	
n	f1	std		f1	std
1	8.2	±1.6	1	**46.8**	-
2	8.9	±2.9	2	**50.8**	-
4	18.1	±3.9	4	51.3	-
8	44.8	±7.1	8	55.0	-
10	48.3	±1.3	10	52.4	-
50	69.2	±2.0	50	61.3	-
100	72.5	±2.6	100	63.4	-
full	**74.1**	±2.5	full	63.4	-
	TARS			**PET**	
n	mean	std		mean	std
1	27.2	±2.1	1	42.8	±5.0
2	33.5	±2.8	2	42.4	±7.6
4	32.8	±3.0	4	**55.8**	±3.9
8	45.9	±3.7	8	**67.6**	±1.7
10	50.0	±3.3	10	**67.3**	±3.7
50	67.3	±1.6	50	**71.9**	±1.4
100	71.9	±0.9	100	**72.9**	±1.6
full	72.4	±1.4	full	73.3	±1.4

Table 4: Few-shot learning performance by macro-F1 (in %; bold: best result for each training set size n)

5. Application

To learn about the potentials and challenges of supervised ABAM for monitoring public media discourses, we apply the PET model in a genre-transfer scenario from web sources in the AAC-NE dataset to a newspaper corpus covering the topic of nuclear energy.

Dataset: Our application scenario focuses on the development of the nuclear energy debate over time. For this investigation, we retrieved a corpus of 3268 newspaper articles from "The Guardian" via the publisher's

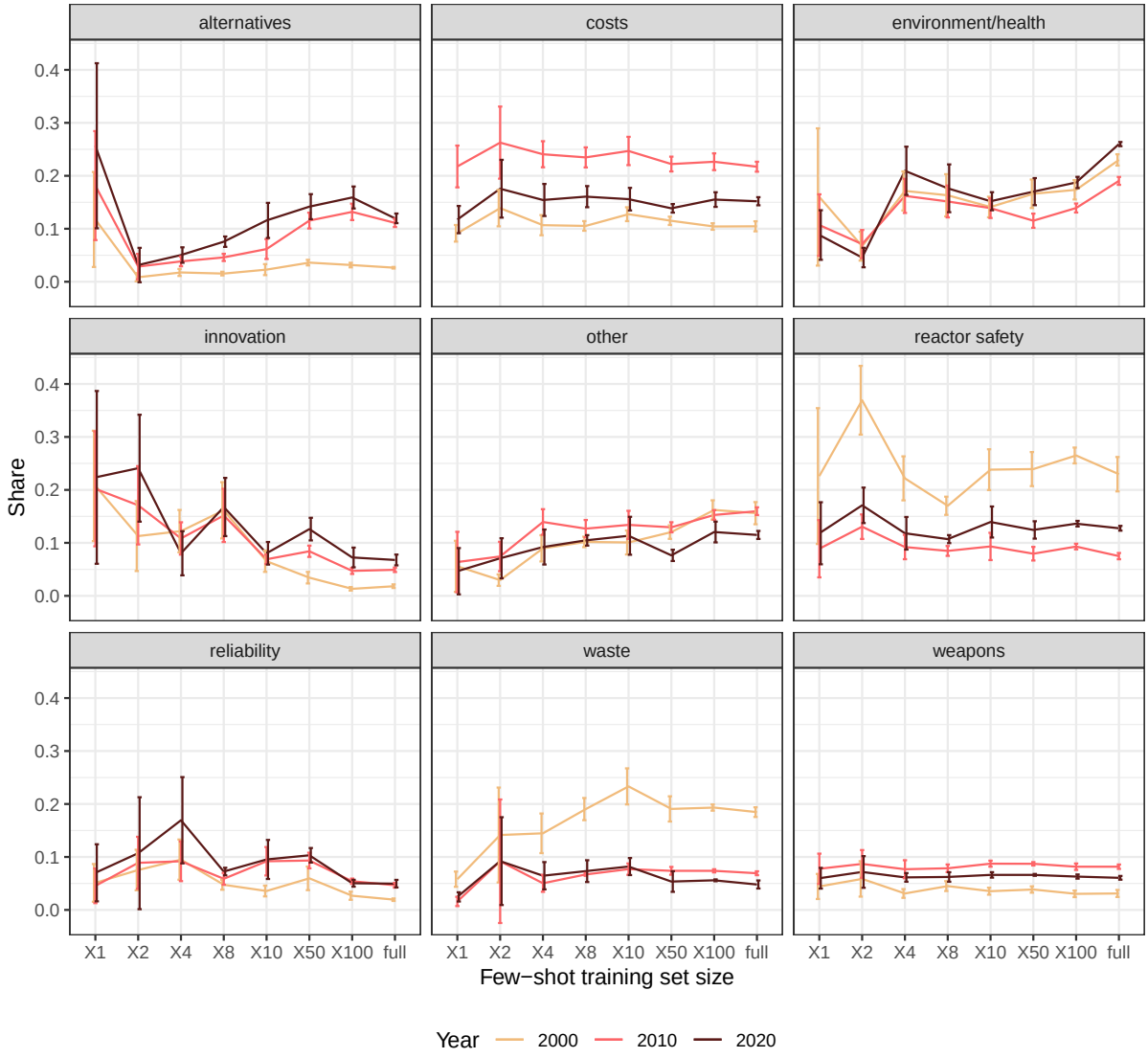

Figure 1: Predicted shares of argument aspects in "The Guardian" with regard to different training set sizes (error bars indicate the standard deviation).

Open Platform[3] API with the R-package *guardianapi* (Odell, 2019). We selected all articles that have been published between 2000 and 2021 and were tagged with "environment/nuclearpower". For argument mining, the articles were split into a total of 97,276 sentences. Further, we fine-tuned a RoBERTa-large model on the UKP SAM dataset (Reimers et al., 2019) to receive 'pro', 'contra', and 'no argument' labels for all sentences.[4] On the UKP-SAM test data, the model achieves a macro-F1-score of 76.1 % across all topics, and 68.1 % for the nuclear energy topic. On the newspaper dataset, the fine-tuned model predicts 26,571 argumentative sentences, 16,902 with a contra stance and 9,669 with a pro stance. In the last step, we feed these argumentative sentences to the 40 PET models (5 repeated runs for 8 training set sizes) to obtain 40 aspect label predictions.

Stability of few-shot predictions: Experiment 2 showed that acceptable performance in terms of F1 scores could be achieved by few-shot learning on in-genre data already with rather small training sets. Due to a lack of true aspect labels, we cannot evaluate the classifier performance in the newspaper genre directly. However, due to the repeated runs of the training with different training data samples of size k, we can observe the stability of predicted aspect shares over time. Stable predictions would indicate that the classifier has been provided with enough information to perform the aspect detection task reliably such that the addition of new information from more training examples does not drastically change the outcome.

[3] https://open-platform.theguardian.com

[4] As input, we concatenate the topic information and the training sentence with a [SEP] token. Hyperparameters are set analogue to Experiment 1.

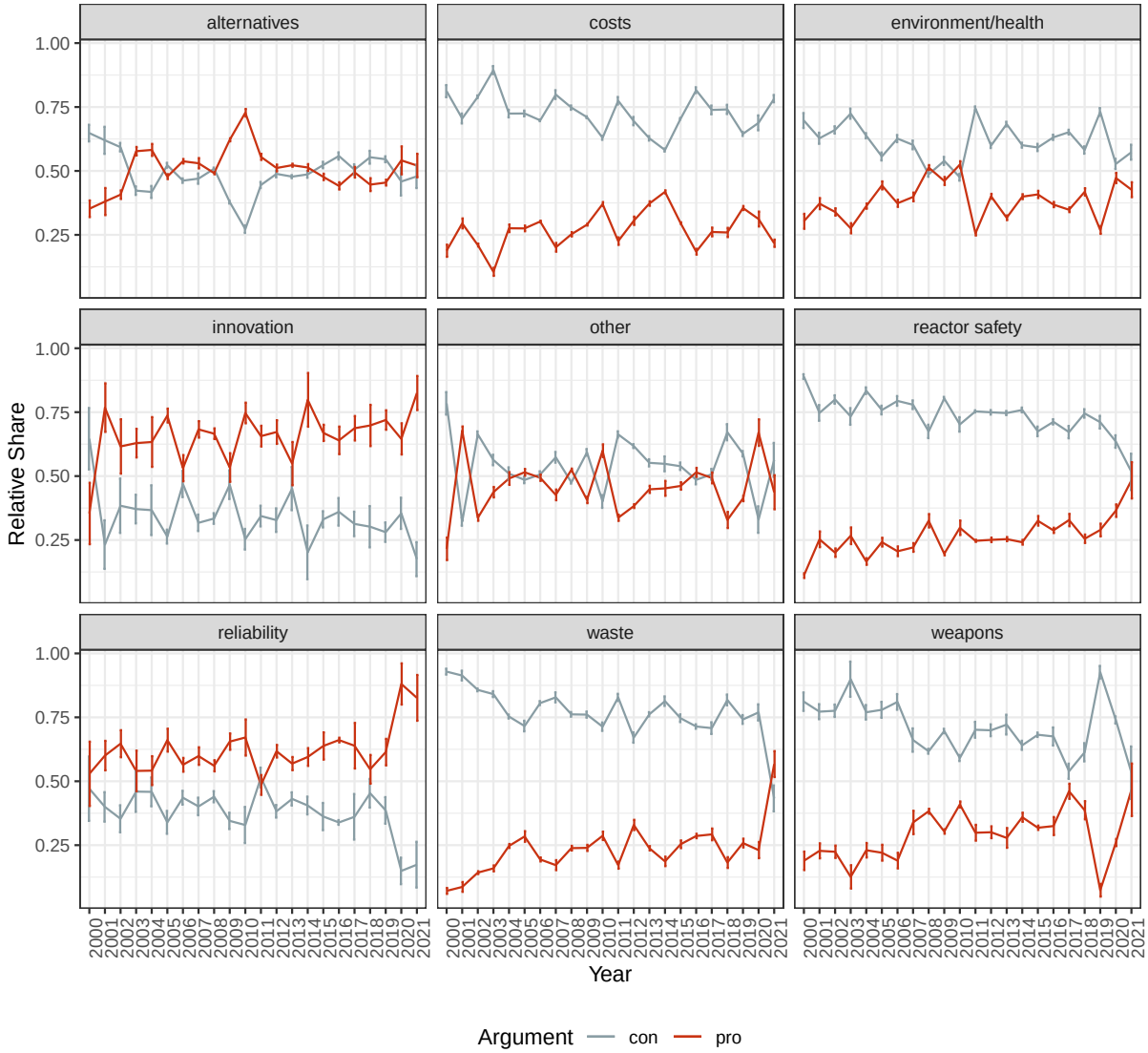

Figure 2: Relative share of pro and contra arguments about certain aspects of the nuclear energy debate in "The Guardian" news articles over time (error bars indicate standard deviation based on predictions from five repeated model inferences).

Fig. 1 shows the stability of predictions at different training set sizes for all aspects in three selected publication years. In general, training based on four or fewer examples leads to an unacceptably large variance in the prediction of shares of single aspects (standard deviation up to ±18 percentage points). For training on 8 and 10 samples, we observe single aspect categories and publication years with stable predictions. But, only for 50 and more samples, the shares per year are predicted stably for all aspects (standard deviation around ±2 to ±4 percentage points). We thus conclude that despite the promising results of few-shot learning on in-sample datasets, we still need medium-sized training sets for robust results in genre-transfer scenarios.

Changing patterns of the nuclear energy debate: Finally yet importantly, we are interested in how aspects relate to pro and contra arguments over time. Fig. 2 shows the relative share of pro/con arguments per aspect over 20 years of newspaper coverage in "The Guardian" as predicted by the five PET models trained on $n = 100$ samples. In general, we see that the aspects *costs*, *environment and health*, *reactor safety*, *nuclear waste* and *weapons* form clear contra arguments mostly. *Technological innovation* and *energy supply reliability*, in contrast, make up for the majority of pro-argument aspects.

Most striking is the development in recent years: although it might be too early to speak of a stable trend, we can observe that pro and con shares of the formerly clearly contra arguments *environment and health*, and *reactor safety* converge. Only the cost aspect remains strongly on the contra-side. In contrast, *innovation* and *reliability* seem to significantly strengthen the pro-stance. These observations point to a clear shift in the

public discourse in favor of nuclear energy in the wake of necessary adaptations of energy policy in the context of climate change.

6. Discussion and Conclusion

In this paper, we introduced the task of supervised aspect-based argument mining along with the AAC-NE dataset (Jurkschat et al., 2022), a subset of the UKP SAM corpus with argument aspect annotations.[5] We show that argument aspects can be annotated and machine-classified on the level of argumentative sentences with sufficient quality for analyzing public media discourses (up to 74 % F1-score, and 78 % accuracy). With the help of few-shot learning techniques based on transformer neural networks,[6] the annotation effort to obtain stable aspect classifications for large, longitudinal news article collections can be reduced to a manageable level. In comparison with the fine-tuned baseline model, few-shot learning approaches provide better results for medium-sized training sets. For very small training set sizes, however, the performance of few-shot learning is not sufficient for our intended purpose. To stably predict argument aspect shares in time slices of a longitudinal dataset, we identified 50 to 100 examples per category as a minimum.

The procedure of detecting, defining, and labeling aspect categories is complex, labour-intensive and opens up several options for improvement. Concerning the manual annotations of the UKP SAM corpus as the basis of our aspect labeling, some arguments are just expressing personal opinions but contain no specific aspect. Further, it contains single sentences only. In media debates, nonetheless, speakers often use multiple sentences to clearly express their argument. In order to facilitate the definition, labeling, and detection of aspect categories, our procedure can be adapted to the requirements of the respective applications in media studies, aiming at changing the level of the coding units from sentences to paragraphs or shorter token sequences, as well as increasing the inter-coder reliability through improved definitions and coder training.

Part of the future work will also be to work on the topic/domain-transfer scenario to further improve few-shot classification. In this study, we applied the best performing aspect classifier, the PET model, in a genre-transfer scenario from web sources in the UKP SAM corpus to a British newspaper corpus, both covering the same topic. Further investigations will focus on the transfer of aspect detection between different argumentative topics to learn how cross-topic transfer of knowledge can improve automatic aspect detection.

In an exemplary application of our approach to a diachronic newspaper dataset, we identified signals for a clear shift in the public discourse in favor of nuclear energy. This revelation of changing patterns of pro and contra arguments related to certain aspects over time showcases the potential of supervised argument aspect detection for tracking issue-specific media discourses. Our workflow can be easily adapted to create corpora and machine classifiers for further debate issues and other languages. It will be part of our future work, to create and publish such datasets for supervised ABAM to perform comparative media studies in the context of computational social and communication science.

[5]The AAC-NE dataset can be downloaded at `https://doi.org/10.5281/zenodo.6470232`

[6]The code for the experiments in this paper is available at `https://github.com/Leibniz-HBI/AAC-NE_experiments`

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG) as part of the project "FAME: A framework for argument mining and evaluation" (project no. 406289255). This work was also supported by the German Federal Ministry of Education and Research (BMBF, 01/S18026A-F) by funding the competence center for Big Data and AI "ScaDS.AI Dresden/Leipzig". The authors gratefully acknowledge the GWK support for funding this project by providing computing time through the Center for Information Services and HPC (ZIH) at TU Dresden.

Bibliographical References

Ajjour, Y., Alshomary, M., Wachsmuth, H., and Stein, B. (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.

Baumann, R., Wiedemann, G., Heinrich, M., Hakimi, A. D., and Heyer, G. (2020). The road map to FAME: A framework for mining and formal evaluation of arguments. *Datenbank-Spektrum*, 20(2):107–113.

Boydstun, A. E., Card, D., Gross, J., Resnick, P., and Smith, N. A. (2014). Tracking the development of media frames within and across policy issues.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations (ICLR)*, pages 26–30, Addis Ababa, Ethiopia. OpenReview.net.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Daxenberger, J., Schiller, B., Stahlhut, C., Kaiser, E., and Gurevych, I. (2020). ArgumenText: Argument

classification and clustering in a generalized search scenario. *Datenbank-Spektrum*, 20(2):115–121.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fraj, M., Hajkacem, M. A. B., and Essoussi, N. (2019). Ensemble method for multi-view text clustering. In *Computational Collective Intelligence*, pages 219–231. Springer International Publishing.

Halder, K., Akbik, A., Krapac, J., and Vollgraf, R. (2020). Task-aware representation of sentences for generic text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3202–3213, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations (ICLR)*, pages 26–30, Addis Ababa, Ethiopia. OpenReview.net.

Lapesa, G., Blessing, A., Blokker, N., Dayanik, E., Haunss, S., Kuhn, J., and Padó, S. (2020). DEbateNet-mig15:tracing the 2015 immigration debate in Germany over time. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 919–927, Marseille, France. European Language Resources Association.

Lawrence, J. and Reed, C. (2020). Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. cite arxiv:1907.11692.

Misra, A., Ecker, B., and Walker, M. (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.

Moens, M.-F. (2018). Argumentation mining: How can a machine acquire common sense and world knowledge? *Argument & Computation*, 9(1):1–14.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Reimers, N. and Gurevych, I. (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.

Schick, T. and Schütze, H. (2021). Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

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

Language Resource References

Jurkschat, L., Wiedemann, G., Heinrich, M., Ruckdeschel, M., and Torge, S. (2022). Argument Aspect Corpus–Nuclear Energy. Distributed via Zenodo, DOI: `10.5281/zenodo.6470232`.

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., and Liang, X. (2018). doccano: Text annotation tool for human.

Odell, E., (2019). *guardianapi: Access the 'Guardian' newspaper open data API*. R package version 0.1.1.

Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., and Gurevych, I. (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.

Stab, C., Miller, T., Schiller, B., Rai, P., and Gurevych, I. (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.

Appendix

Aspect	Description
alternatives	Other ways for energy supply like renewable energy sources or coal.
costs	Costs for producing nuclear energy, maintaining nuclear power plants or building them.
environment/heath	Impacts on nature and human beeings. Health risks and pollution.
innovation	Improvement in the production of nuclear energy e.g. technically or environmentally.
reactor safety	Technical security of nuclear power plants or reactor disasters (failed security).
reliability	Amount of generated power and the supply trustworthiness.
waste	The waste problem that is incurred when producing nuclear energy.
weapons	Usage of nuclear energy for any kind of weapons or war or terrorism.
other	Arguments that do not belong to any aspects above.

Table A1: AAC-NE label definitions

Aspect	Precision	Recall	F1-score
alternatives	60.7	81.0	69.4
costs	80.0	82.8	81.4
environment/health	79.7	85.9	82.7
innovation	50.0	25.0	33.3
other	75.0	62.1	67.9
reactor safety	84.2	74.4	79.0
reliability	55.6	50.0	52.6
waste	85.2	88.5	86.8
weapons	87.5	93.3	90.3

Table A2: Classification performance for individual aspects (RoBERTa model fine-tuned on the full training set)