Boundary Detection and Categorization of Argument Aspects via Supervised Learning

Mattes Ruckdeschel and **Gregor Wiedemann**
Leibniz-Institute for Media Research | Hans-Bredow-Institute, Germany
{m.ruckdeschel, g.wiedemann}@leibniz-hbi.de

Abstract

Aspect-based argument mining (ABAM) is the task of automatic *detection* and *categorization* of argument aspects, i.e. the parts of an argumentative text that contain the issue-specific key rationale for its conclusion. From empirical data, overlapping but not congruent sets of aspect categories can be derived for different topics. So far, two supervised approaches to detect aspect boundaries, and a smaller number of unsupervised clustering approaches categorizing groups of similar aspects have been proposed. In this paper, we introduce the Argument Aspect Corpus (AAC) which contains token-level annotations of aspects in 3,547 argumentative sentences from three highly debated topics. This dataset enables both the supervised learning of boundaries and the categorization of argument aspects. During the design of our annotation process, we noticed that it is not clear from the outset at which contextual unit aspects should be coded. We, thus, experiment with classification at the token, chunk, and sentence level granularity. Our finding is that the chunk level provides the most useful information for applications. At the same time, it produces the best-performing results in our tested supervised learning setups.

1 How to Code Argument Aspects?

Argument mining has become a prominent natural language processing task with several challenging sub-tasks (Lawrence and Reed, 2020). Argumentative utterances are found plentiful in online forums, newspapers, and social media debates, which offer heaps of text data for argument mining. Depending on the variety and complexity of the issues of a given topic, the number of talking points in such debates could be potentially very large. However, with the concept of theoretical or thematic 'saturation,' qualitative researchers refer to the fact that public debates typically revolve around a relatively small set of issues that can be inferred from textual data with manageable manual effort (Johnson, 2014). These issues are accompanied by a likewise limited set of prototypical arguments. To describe the width and depth of a debate on a given topic, arguments can be grouped according to their reference to the same aspects. Analog to Schiller et al. (2021), we define an aspect as a semantically distinguishable, recurring subtopic of an argument that expresses the issue-specific key rationale for its conclusion. A stance on an aspect, thus, potentially serves as a justification for the stance on the corresponding main topic that itself can but not necessarily has to be mentioned in the argument.

For example, in the argument *"Businesses are sometimes forced to [hire fewer employees] because they must pay minimum wage"* the token sequence in brackets holds the key rationale for the aspect category *(un-)employment rate*. In contrast, in a slightly modified version of this argument *"[Businesses were sometimes forced to close down] because they must pay minimum wage"* the sequence in brackets refers to the aspect category *competition/business challenges*. Both argument versions implicitly express a negative stance on statutory minimum wages as higher unemployment or increased bankruptcies of businesses are generally seen as undesired policy outcomes.

Individual arguments may refer to different aspects that perhaps even take opposing stances before giving reason for a final stance. Extracting aspects from arguments has several advantages for the analysis of debates in various disciplinary settings such as political science, social science, or economics. *First*, it adds a new semantic dimension to the established identification of structural components in argument mining. This allows for a theory-led grouping of relevant talking points that can facilitate a qualitative discourse inspection. For quantitative analysis, they allow for investigating the prevalence of aspects in specific debates and their co-occurrence with argumentative stances as well as other aspects. *Second*, aspects as semantic

Proceedings of the 9th Workshop on Argument Mining, co-located with the 29th International Conference on Computational Linguistics, pages 126–136
October 17, 2022.

categories can serve as a bridge to combining argument mining with the formal modeling of argument semantics (Baumann et al., 2020).

Often, aspects are neither explicitly stated nor consistently formulated in arguments which makes unsupervised aspect category extraction practically infeasible. Instead, we argue for the creation of well-defined and systematically controlled aspect category sets that generalize key points in similar arguments against the background of domain knowledge to serve the purpose of ABAM. This abstracts from the complexity and diversity of aspect expressions so that only a limited number of aspect categories are required to fully cover a topic. This not only enables manual coding and supervised classification, but guarantees a methodologically and theoretically sound interpretation of the classification results. It further enables comparative studies across divergent datasets that can hardly be achieved solely by relying on unsupervised methods.

To perform supervised ABAM, we created the Argument Aspect Corpus—a data set for supervised learning of aspects for three topics. In this paper, we describe the iterative process for creating aspect categories for a given topic, starting from an unsupervised clustering of arguments and refining aspect categories after coding samples from a data set in several rounds. During the design of our annotation process, we realized that it is not clear from the outset at which contextual unit aspects should be coded. We started with a multi-label sentence classification task but soon noticed that confining the label decision to a certain token sequence within a sentence not only would provide more valuable information for aspect mining, but also leads to better justified and, thus, more coherent label decisions. However, for a sequence tagging task, unlike for named entity recognition, span boundaries are much less obvious. If the annotated span is too wide it may contain unnecessary information to capture the aspect and, thus, distract a machine learning process from the actual task. If the span is too small, the annotated text may not represent the aspect properly.

In light of these considerations, we answer the research question: What is the recommended level of granularity to perform supervised ABAM? Hence, there are two main contributions of our paper:

1. We introduce the *Argument Aspect Corpus* (AAC) for supervised aspect-based argument mining. It contains 3,547 argumentative sentences from three highly debated topics: nuclear energy, minimum wages, and marijuana legalization.
2. We perform experiments to determine the optimal granularity of aspect boundaries. For this, we test token-based and chunk-based multi-class classification against multi-label sentence classification for argument aspects. We identify a sequence tagging task based on chunk-normalized tokens as the recommended approach.

In Section 2 we relate our approach to ABAM to several other approaches for the semantic grouping of arguments. We then present our data sets and explain our iterative annotation process in Section 3. Section 4, describes our experiment setup and the reasoning behind it. Section 5 describes our experiments on the automatic prediction of aspect labels with state-of-the-art transformer networks, as well as the optimal aspect granularity. We will present the main findings and conclusions of our work in section 6.

2 Related Work

During the last years, several approaches to grouping arguments into some type of semantic categories were published in the field of argument mining. To describe their task, these approaches rely on heterogeneous names, theoretical concepts, and mining strategies. A first group of approaches builds on *framing* theory that is commonly used in empirical communication and media research. In argument mining, the notion of a frame is adopted as the aspect of a discussion that is emphasized by an argument. Sets of aspects can be of varying breadth and depth. Also, approaches differ whether they assume frames to be issue-specific or should generalize across topics. Ajjour et al. (2019), for instance, define a *frame* as a set of arguments that focus on the same aspect. To identify references to the same aspects, they use an unsupervised clustering on argumentative texts. By definition then, each cluster supposedly represents one frame. However, the resulting clusters do not necessarily describe semantic frames in the sense of repeatedly occurring aspects of the corresponding discussion. The large number of optimal clusters as described in the paper also drastically reduces the usefulness for any further study. Heinisch and Cimiano (2021) define frames as the aspects a talking point discusses. They address the shortcomings of frames that are too generic and frames that are too issue-specific by clustering user-generated, specific labels into general frame categories from classic media research. Although they have shown that their approach is able to automatically identify media frames to some extent, they do not provide well-defined sets of issue-specific aspects that would allow for a deductive analysis of public debates. Daxenberger et al. (2020) describe a clustering-based grouping of arguments based on aspects for better search results. They use agglomerative hierarchical clustering of contextualized word embeddings, such as BERT-embeddings (Devlin et al., 2019), on sentence-level argument pairs. The resulting clusters based on similarity metrics also do not necessarily provide useful aspect categories, let alone semantically meaningful labels.

Bar-Haim et al. (2020) introduced *key point analysis* to generate a summary for large collections of arguments by finding *key points*. Their work also inspired the ArgMining 2021 shared task (Friedman et al., 2021) which contained one task for matching arguments to key points, and one task for the generation of key points. Hereby, key points are defined as higher-level arguments that occur frequently in debates on a given topic. Key points are formulated as full sentences and with an indication of a clear pro or contra stance on the debated issue. Besides the difference that in our definition aspects are independent of any stance, key points can play a similar role in argument classification as our proposed aspects. They also acknowledge the difficulty of the problem of argument grouping and the ineffectiveness of unsupervised methods based on contextual embeddings.

Addressing the problem of unsupervised approaches, Jurkschat et al. (2022) propose ABAM as a multi-class sentence classification task and provide a corpus containing argumentative sentences from the nuclear energy debate with manually annotated class labels. In a further development of this work, our approach to ABAM is designed as a token-level sequence tagging task that allows for multiple aspects to being mentioned in one sentence, and for the extraction of the decisive sentence parts determining these aspects.

Annotating and predicting aspects on the token level is also performed in the works of Trautmann (2020) and Schiller et al. (2021). Trautmann (2020) defines *aspects* analog to aspect-based sentiment analysis (Pontiki et al., 2016). He proposes the task of *Aspect Term Extraction (ATE)* and presents a supervised sequence tagging approach to detect the most common token n-grams that address argument aspects. However, no semantically meaningful aspect categories are created from the extracted token sequences. Similar to ATE, Schiller et al. (2021) perform aspect boundary detection as a supervised sequence tagging task trained on argumentative sentences in which token sequences were labeled with a BIO-tagging scheme to indicate the beginning (B), inside (I) and outside (O) of token aspect spans. They also address the problem of fuzzy span boundaries that motivated our research and present a crowdsourcing task based on automatic candidate ranking and manual candidate selection to create a gold standard with high inter-coder agreement. Regarding this task of aspect boundary detection, their approach to ABAM mostly resembles ours. We, however, extend the tagging and extraction of aspect terms to a classification of the predicted spans into issue-specific aspect categories.

3 The Argument Aspect Corpus

With this paper, we publish the Argument Aspect Corpus (AAC) that contains manually annotated aspect labels on token spans from argumentative sentences. The argumentative sentences were extracted from the UKP Sentential Argument Mining Corpus (UKP SAM) (Reimers et al., 2019). For the AAC, we selected only those sentences that have been annotated as either expressing a pro or a contra stance on one of the three topics: *minimum wage (MW)*, *nuclear energy (NE)*, and *marijuana legalization (MJ)*. The topics were chosen with respect to their importance within recent European political discourses.

As Bar-Haim et al. (2020) and Jurkschat et al. (2022) have already pointed out, labeling aspects in arguments is a complex task. This is mainly due to the fact that the granularity of aspects cannot be determined in a data-driven manner, but must be specified in a methodically rigorous process of developing the coding scheme. With this comes the necessity to develop definitions of aspect categories that are as precise as possible to separate the sometimes overlapping meanings of argumentative components from one another. To fulfill these requirements and, at the same time, address the heterogeneity of the empirical data, we followed a process that combined unsupervised clustering with group discussions to reach consensus definitions of our aspect categories. As a starting point, we employed unsupervised k-means clustering of sentence embeddings from S-BERT (Reimers and Gurevych, 2019). Analog to previous research on argument frames, we expect that semantic-similarity-based clusters already group aspect information to some extent. We decided on a fixed number of 15 initial clusters as a rough estimate of how many aspects per topic we expect. However, our subsequent development of aspect categories would allow for the creation of more or fewer aspect categories. With a group of three annotators, students, and researchers from the field of (computational) social science, we listed aspects that occur in these clusters as a first summary of a topic. With these initial aspects, we created a preliminary codebook and annotated a sample of 200 sentences per topic. Arguments in these samples were sorted by cosine similarity of their S-BERT representation. Annotators reported that this sorting was beneficial for speeding up the annotation and, at the same time, increasing its coherence. Annotators were encouraged to write comments about aspect categories and extend the list of aspects if necessary. Next, the inter-coder agreement (Krippendorff's alpha) for each aspect was calculated on a sentence level in order to find aspects that need clarification. In extensive discussion rounds, the category definitions were sharpened and refined. In the second and following rounds of annotating samples, we switched from the sentence level classification to an annotation of token spans to be able 1) to justify label decisions directly on text snippets, and 2) to allow for aspect term extraction in a subsequent step of machine learning. This iterative process of annotation, agreement evaluation, and discussion was repeated until a consensus for all aspect definitions was reached and the list of aspects covered the large majority of arguments for a topic. The full dataset was then annotated by all three annotators with the final codebook resulting from the aforementioned iterative process. A major challenge during annotation was determining token span boundaries since in many cases it is not possible to unambiguously decide where the mentioning of an aspect in a sentence actually starts or ends. We decided to instruct annotators to label the smallest number of tokens that provide sufficient information to label the aspect on its own. Still, this resulted in substantial disagreement about aspect boundaries in many cases while, at the same time, sentence-level agreement of labeled aspects was high. This observation led to the decision to further investigate the question of which granularity level of context units ABAM should be performed (cp. Section 5).

Final gold labels for the AAC dataset on the token level were derived in a two-step process. First, on the sentence level, we determined all labels that have been annotated by a majority of annotators as gold labels. Arguments without any majority label were reviewed once again to determine a final label. Second, for each token in a sentence, we copied the sentence gold label if at least one annotator included it in his/her annotation span. Again, rare conflicts of overlaps of sentence majority labels for individual tokens have been resolved in a final review. This strategy results in potentially more extensive gold labels on the token level compared to those of the single annotators.

Table 1 provides an overview of the dataset statistics of the AAC. Due to the challenge of achieving exact matches on span boundaries during the annotation, we opted for an inter-coder agreement measure on the sentence level. For each topic, this was calculated using Krippendorff's alpha in combination with the MASI distance (Passonneau, 2006) as a weighted agreement metric over the set of all labels that an annotator has used to label a sequence in a sentence. Thus, only if two annotators use the exact same set of labels to annotate a sentence, the resulting distance is 0. With alpha values of 0.65 and higher, we achieve acceptable agreement between coders. But the numbers also signal that argument aspect coding is a challenging task that requires a certain amount of coder training and expertise. Measures of the agreement for individual aspect categories some of which are significantly higher than the overall agreement are reported in Tables 7, 8, and 9 in the Appendix.

4 Experimenting with Aspect Boundaries

In a significant number of cases, annotators agreed upon which aspects were present in an argument but labeled slightly different token sequences as indicative of an aspect. Therefore, the strict token-level annotator agreement was relatively low compared to the agreement on the sentence level. A qualitative look into boundary disagreement for a small sample revealed that different individual an-

Topic	Aspects	N	α_K	Aspect categories
Minimum Wage (*MW*)	12	1118	0.65	motivation/chances, competition/business challenges, prices, social justice, welfare, economic impact, turnover, capital vs. labour, government intervention, un/employment rate, low-skilled and secondary wage earners
Nuclear Energy (*NE*)	12	1261	0.68	waste, accidents/security, reliability, costs, weapons, technological innovation, environmental impact, health effects, renewables, fossil fuels, energy policy, public debate
Marijuana Legalization (*MJ*)	13	1213	0.65	child and teen safety, community/societal effects, health/psychological effects, medical marijuana, drug abuse, illegal trade, personal freedom, national budget, drug policy, addiction, harm, gateway drug, legal drugs

Table 1: AAC Dataset statistics: the number of aspects, the number of arguments (N), Krippendorff's inter-coder agreement (α_K) and the aspect categories for all three topics of the current version.

notations could be considered valid regarding our guidelines. This challenge to achieve a high agreement for exact matches of token span boundaries during aspect annotation led us to the more general questions: what would be the most suitable level of granularity of context units, and what would be the best corresponding modeling approach to perform ABAM as a machine learning task?

To answer these questions, we experiment with different modifications of the AAC dataset. Since the category *Other* was used to annotate any sentence that either did not fit any aspect definition or was deemed not argumentative, we excluded the category from training. Then, we split the annotated data per topic randomly into a training (70 %), validation (10 %), and test set (20 %), Finally, we created different formats of these sets to test different ABAM task variants:

- **Sequence tagging:** Analog to named entity recognition (NER), each token is labeled either with its gold aspect category or the `O`-tag. Unlike Schiller et al. (2021), we refrained from using BIO(ES) prefixes to indicate beginning, inside, end, or single-token tags during training since our annotation guidelines do not allow adjacent sequences of distinct aspects of the same category. We further noticed during early experiments that BIO-tags significantly harmed the overall performance. With this input, we fine-tune a pre-trained transformer model with a sequence tagging head.[1]

- **Chunk normalization:** To improve the coherence of aspect boundaries within the dataset, we utilized information from a syntactic chunker.[2] We hypothesize that syntactic chunks are a more suitable level of context compared to sentences and tokens. They are more fine-grained than sentence-level annotations but more coarse-grained and, thus, coherent for machine learning and prediction than token-level annotations. Chunk normalization is performed by copying aspect labels from each annotated token to all other tokens belonging to the same chunk.

- **Multi-class chunk classification:** In this variant of the task, we do not strive for the prediction of labels of individual tokens but entire chunks. For this, we feed each target chunk and its surrounding sentence separated with a `[SEP]` token into a transformer model with a final multi-class output layer. Gold chunk labels are derived from the AAC gold labels the same way as for the chunk normalization.

- **Multi-label sentence classification:** High levels of inter-coder agreement on the sentence level might also suggest that ABAM is performed best as a sequence classification task for argumentative sentences neglecting aspect spans. In contrast to chunks, sentences can refer to multiple aspects, thus, requiring a multi-label classification. To test for this simplified version of the task analog to the

[1] All experiments are conducted with the *Flair* NLP framework (Akbik et al., 2019).

[2] We used the pre-trained English chunker model from *Flair* (Akbik et al., 2019).

Task variant	Sentence representation examples
Sequence tagging	[After] [the] [wage] [increase] [,] [that] [same] [basket]$_{\text{prices}}$ [cost]$_{\text{prices}}$ [\$] [315] [.]
Chunk normalization	[After] [the] [wage] [increase] [,] [that]$_{\text{prices}}$ [same]$_{\text{prices}}$ [basket]$_{\text{prices}}$ [cost]$_{\text{prices}}$ [\$] [315] [.]
Chunk classification	[that same basket [SEP] After the wage increase , that same basket cost \$315.]$_{\text{prices}}$
Sentence classification	[After the wage increase , that same basket cost \$315.]$_{\text{prices}}$

Table 2: Examples the four task variants tested for supervised ABAM (brackets indicate context unit boundaries, sub-scripted text indicates the aspect label).

Model	Precision	Recall	F1
	Minimum Wage		
roberta-large	**58.4±1.7**	**75.1±2.0**	**65.7±1.8**
albert-large-v2	35.8±3.8	50.1±4.9	41.7±4.3
electra-large	44.3±10.3	55.9±14.5	49.3±12.1
	Nuclear Energy		
roberta-large	**63.6±0.9**	**78.2±0.7**	**70.1±0.4**
albert-large-v2	51.8±2.9	66.9±3.8	58.3±2.2
electra-large	62.6±1.2	75.8±1.4	68.6±1.1
	Marijuana Legalization		
roberta-large	**60.5±2.4**	**76.8±1.8**	**67.4±1.9**
albert-large-v2	39.5±2.4	58.7±2.5	47.2±2.4
electra-large	42.0±20.8	52.4±23.2	46.2±22.4

Table 3: Performance of token-level aspect tagging for three different topics (metrics in %, entity-type evaluation scheme, mean and standard deviation of five repeated runs).

approach by Jurkschat et al. (2022), we reformat the AAC dataset splits into full sentences with a set of gold labels from all contained tokens to fine-tune a transformer model with a multi-label classification head.

Table 2 shows the differences between the inputs for the two sequence tagging and the two sequence classification tasks. Since the token *basket* was annotated in the AAC gold labels, the entire chunk *that same basket* becomes annotated in chunk-normalization.

5 Supervised ABAM

First, we perform a step of model selection to determine the best pre-trained language model for performing ABAM. Second, we test different modeling variants of the ABAM task to learn about the most fitting context units for argument aspects.

5.1 Language model selection

We test several state-of-the-art language models on the aspect classification tasks in the variant of sequence tagging. We compare three common language models: RoBERTa-large (Liu et al., 2019), ALBERT-large (Lan et al., 2019), and ELECTRA-large (Clark et al., 2020). To ensure the stability of results, all experiments were repeated five times with different random seeds. In our first tests, XLM-RoBERTa (Conneau et al., 2020) performed significantly worse than RoBERTa and was, therefore, excluded from further testing. ALBERT-large was chosen over ALBERT-xxlarge, since the results for the xxlarge model version were not significantly better during first runs, whereas computing time increased significantly. All tests were conducted with the same set of reasonable default hyper-parameters (see Table 11 in the Appendix).

Table 3 shows the performance of the tested language models which were obtained using the *entity type* evaluation scheme of the *nervaluate*[3] python package.[4] Since the annotation of aspect boundaries was somewhat incoherent for individual arguments between multiple annotators, we expect coherency also to be affected across different arguments within the AAC gold annotations. For this reason, the entity type evaluation scheme appears as the right choice, because instead of exact span boundaries it considers overlapping of predicted and gold spans to be a correct prediction, as long as the annotated labels of the overlapping spans match.

With F1-scores between 65.7 % and 70.1 %, the RoBERTa model outperforms the other models on the task significantly.[5] Therefore we decided to continue granularity experiments only for

[3] https://pypi.org/project/nervaluate/

[4] *nervaluate* implements different evaluation schemes for sequence tagging based on Segura-Bedmar et al. (2013).

[5] The lower performance of the other models can be explained by the fact that they completely failed to predict some smaller aspect categories. We hypothesize that an extended search for more optimal hyper-parameters would lead to considerable performance increases. Since this is not the main focus of this paper, we decided to leave this for future work.

RoBERTa-large. We also observe that the recall is consistently and significantly higher than precision.

5.2 Aspect granularity evaluation

To test different variants of modeling the ABAM task, we fine-tune a RoBERTa-large model for each topic of the AAC dataset separately. To make the results of these variants comparable, we convert the predictions of all models to the coarsest granularity of sentence-level aspect labels. We compare the set of labels that were predicted for all tokens or chunks of a sentence to the set of gold standard sentence labels. Table 4 shows the micro-average performance of the various models.[6]

With F1-scores of 80.2% and higher, all models that classify aspects finer than sentence level granularity achieve not only very satisfactory results, but also significantly outperform aspect mining on the sentence level. This is a clear hint that ABAM profits from finer-grained annotation levels. The results also show that sentence-level classification achieves the best precision values, but suffers from lowered recall. This shows that labeling on the token or chunk level can provide more valuable and consistent insight into the used aspects in a sentence or argument. Sentence-level aspect classification, in contrast, often seems to overlook aspects that differ too much from the training sentences. Normalizing token-level annotations to chunk boundaries slightly improves the recall and accuracy compared to basic sequence tagging for the topics of minimum wage and nuclear energy. For the other metrics, the effect is ambiguous.[7] We conclude that chunk normalization may be useful to make annotation spans more consistent and therefore improve classification results slightly, although the effect is not large. Models trained to classify chunks along with their sentence context directly perform consistently worse compared to models trained on token-level sequence tagging.

5.3 Multi-topic aspect classification

In the last experiment, we want to find out whether combining data from several topics produces superior models for aspect classification compared to models trained on a single topic. As a basis, we use the chunk-normalized token dataset. Each argument token sequence is extended by preceding it with tokens containing their respective topic name followed by a separator token ([SEP]) (for an example, see Table 10 in the Appendix). Table 5 shows the performance of the trained multi-topic model over all three topics and the corresponding performance improvement compared to the single-topic classifiers. All topics benefit from the additional training data from other topics. The F1-scores improved significantly up to $+5.7\ \%$. The improvements in precision are considerably higher than for recall. The results show that more training data can improve model performance, even in a multi-topic setting. It is notable, that the improvement for the dataset about nuclear energy has the lowest improvement while being the dataset with the highest inter-coder agreement. This suggests that the multi-topic classifier was able to enhance the results of the slightly less coherently labeled datasets even further.

5.4 Error analysis

To learn about common error patterns, we take a closer, qualitative look at samples of false positives and false negatives of predicted aspect sequences, as well as wrongly classified aspect categories. Table 6 shows three example arguments from the minimum wage topic with aspect labels as predicted by our best-performing single-topic classifier.

In the first example, the model predicted additional spans for the same aspect (false positives). On closer inspection, these annotations can also be considered valid suggesting that the gold annotations are not entirely consistent. Annotating a large dataset with multiple annotators consistently is challenging. This is especially true for complex and potentially overlapping categories such as argument aspects. The example also supports the impression that for real application scenarios the precision values may indicate lower than actual model quality. The second example shows a minimal annotation span by the model that misses the wider span boundaries from the gold standard (false negatives). Here, the model was not able to see the same connectivity between *keep wages down* and *and keep unions out*, which was more apparent to a human annotator. Nonetheless, the model predicted the correct label for the correctly identified aspect token which makes the result partially useful for application scenarios. The last example

[6] Higher values of the F1-score compared to accuracy originate from the span-based evaluation with `nervaluate` compared to the token-wise evaluation for accuracy.

[7] A positive effect from chunk normalization on the results up to +3 percentage points can be observed when using the `strict` evaluation scheme of nereval that compares sequences of exact matches between predicted and gold labels.

Task variant	Precision	Recall	F1	Accuracy
Minimum Wage				
Sequence tagging	77.1±1.2	84.1±2.4	80.4±1.4	66.0±2.0
Chunk normalization	77.1±0.5	**84.7±2.1**	**80.7±1.0**	**66.0±1.0**
Chunk classification	74.9±1.4	86.2±0.9	80.2±0.7	64.3±2.5
Sentence classification	**84.3±1.4**	67.9±1.2	75.2±1.3	64.3±1.2
Nuclear Energy				
Sequence tagging	77.9±0.7	88.0±1.0	**82.6±0.6**	63.7±1.0
Chunk normalization	75.6±1.0	**88.5±2.4**	81.5±1.3	**65.7±1.9**
Chunk classification	74.4±2.3	87.8±1.3	80.5±1.1	61.9±1.9
Sentence classification	**83.8±0.9**	62.4±0.6	71.5±0.7	60.3±0.7
Marijuana Legalization				
Sequence tagging	79.4±1.3	87.1±1.8	**83.1±1.4**	**70.0±2.3**
Chunk normalization	78.0±1.6	**87.0±1.2**	82.3±1.1	68.1±1.2
Chunk classification	76.9±1.4	88.6±1.5	82.3±0.6	66.6±1.6
Sentence classification	**82.2±2.3**	68.5±1.9	73.8±2.1	68.8±1.4

Table 4: Micro-average performance (in %) of four modeling variations of aspect granularity. The test set predictions of the token and chunk-based approaches have been converted to a multi-label sentence prediction to allow for a fair comparison (mean and standard deviation of five repeated runs).

Topic	Precision	Impr.	Recall	Impr.	F1	Impr.
Minimum wage	66.2±2.6	+6.1%	76.9±1.6	+0.9%	71.1±2.1	+3.7%
Nuclear energy	64.7±1.4	+2.4%	80.2±1.5	+3.0%	71.5±1.1	+2.7%
Marijuana legalization	67.6±1.2	+8.8%	80.4±1.3	+2.1%	73.4±0.3	+5.7%

Table 5: Performance of the multi-topic sequence tagging model for argument aspects on chunk-normalized tokens (metrics in %, entity-type evaluation scheme, mean and standard deviation of five repeated runs). *Impr.* is the percentage improvement compared to single-topic models.

shows a wrongly predicted aspect category. The abstract proverb *to move up the economic ladder* was interpreted by annotators to indicate an opportunity for an employee to improve. The model, however, interpreted it as referring to low-skilled workers. This example also shows the difficulty of the task, for humans, and machines. For individual arguments, aspect categories still may have some overlap, even if they were carefully crafted to be about distinct sub-topic of the discourse. Deciding which category is the most suitable becomes even more difficult if metaphorical language is used.

6 Conclusion

In this paper, we further defined the task of supervised aspect-based argument mining based on experiments with a newly created dataset containing aspect annotations of token spans in argumentative sentences from three different topics. With our experiments,[8] we showed that ABAM performs best on a granularity level finer than multi-label sentence classification (cp. Exp. 2). We also showed that best results are achieved by fine-tuning a state-of-the-art language model such as RoBERTa on a token sequence tagging task. Despite satisfactory results up to 70 % F1-score (cp. Exp. 1), we see that especially disagreement on span boundaries for annotated aspects is a source of error. Normalizing token labels in the gold dataset to identical labels within syntactic chunks can mitigate this effect to some extent (cp. Exp. 2). Compared to sentences that can refer to multiple aspects, chunks are short enough to carry information for only one aspect. Compared to tokens, chunks contain more

[8] The AAC dataset and the experiment code for this paper is available at `https://github.com/Leibniz-HBI/argument-aspect-corpus-v1`.

Error type	Argument
False positives	Supporters of minimum wage also believe that a minimum wage stimulates consumption$_{\text{[Economic Impact]}}$ and thus puts more money$_{\text{[Economic Impact]}}$ into the economy$_{\text{[Economic Impact]}}$ by allowing low paid workers to spend more$_{\text{[Economic Impact]}}$.
False negatives	They've been using undocumented immigrants for DECADES (in violation of the law) to keep wages down , and$_{\text{[Capital vs. Labour]}}$ unions$_{\text{[Capital vs. Labour]}}$ out$_{\text{[Capital vs. Labour]}}$.
False category	Minimum wage laws can lead to labor market rigidities$_{\text{[Motivation/Chances]}}$ that make it more difficult for people to move up$_{\text{[Low-skilled]}}$ the economic ladder $_{\text{[Low-skilled]}}$.

Table 6: Examples for false predictions of the best performing aspect classification model (RoBERTa-large, chunk-normalized token sequence tagging). Text color blue indicates true positives, black true negatives. Background colour highlighting indicates errors (green: false positives, gray: false negatives; red: wrong aspect category). For the last example, the correctly identified aspect span was labelled as 'Motivation/Chances' in the gold standard.

information that can be interpreted unambiguously and have clear sequence boundaries that seem to support more consistent manual and automatic data annotations. In addition, the annotation process can be accelerated by tasking annotators with coding chunks instead of sequences or tokens.

In future work, we, therefore, concentrate on a new chunk-based annotation and classification pipeline for ABAM. The results from our third experiment on multi-topic classification will also be of additional help for ABAM research and applications. Training one model on all three topics with a merged set of aspect categories further improved the F1-score of our best model up to 5.7 %. This result is also promising for developing the approach further into a zero-shot or few-shot scenario for yet unseen topics as it was tested successfully already on the sentence level by Jurkschat et al. (2022). With this paper, we publish the Argument Aspect Corpus (AAC) in its version 1.0 containing aspect category definitions, annotation guidelines, and token-level annotated sentences for three topics. Our aim is to provide more topics in future versions, paired with the research about the efficacy of chunk-level annotation processes and few-shot classification performance.

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). The authors gratefully acknowledge the work of their coding team, Aya Alwais and Emma Neuhaus, who coded thousands of sentences and competently supported the definition of aspect categories with their social science domain expertise.

References

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.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. 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.

Ringo Baumann, Gregor Wiedemann, Maximilian Heinrich, Ahmad Dawar Hakimi, and Gerhard Heyer.

2020. The road map to FAME: A framework for mining and formal evaluation of arguments. *Datenbank-Spektrum*, 20(2):107–113.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia. OpenReview.net.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 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.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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.

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.

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

Lauren Johnson. 2014. Adapting and combining constructivist grounded theory and discourse analysis: A practical guide for research. *International Journal of Multiple Research Approaches*, 8(1):100–116.

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 Language Resources and Evaluation Conference*, pages 663–672, Marseille, France. European Language Resources Association.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. Cite arxiv:1907.11692.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation*, pages 19–30.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 567–578. Association for 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.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. 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.

A Appendix

Minimum Wage Aspects	α_K
Un/employment rate	0.80
Motivation/chances	0.67
Competition/business challenges	0.58
Prices	0.88
Social justice/injustice	0.70
Welfare	0.76
Economic impact	0.80
Turnover	0.96
Capital vs labour	0.51
Government	0.65
Low-skilled	0.69
Youth and secondary wage earners	0.58
other	0.56
all topics	0.65

Table 7: Intercoder-agreement for all topics form the minimum wage dataset (Krippendorff-alpha α_K)

Nuclear Energy Aspects	α_K
Waste	0.89
Health effects	0.77
Environmental impact	0.75
Costs	0.79
Weapons	0.88
Reliability	0.59
Technological innovation	0.67
Energy policy	0.66
Renewables	0.94
Fossil fuels	0.89
Accidents/security	0.79
Public debate	0.63
Other	0.64
all topic	0.68

Table 8: Intercoder-agreement for all topics from the nuclear energy dataset (Krippendorff-alpha α_K)

Marijuana Legalization Aspects	α_K
Illegal trade	0.87
Child and teen safety	0.89
Community/Societal effects	0.54
Health/Psychological effects	0.78
Medical Marijuana	0.92
Drug abuse	0.78
Addiction	0.95
Personal freedom	0.79
National budget	0.77
Gateway drug	0.90
Legal drugs	0.91
Drug policy	0.50
Harm	0.53
Other	0.49
all topics	0.64

Table 9: Intercoder-agreement for all topics from the marijuana legalization dataset (Krippendorff-alpha α_K)

Token id	**Text**	**Label**
1	minimum	O
2	wage	O
3	[SEP]	O
4	After	O
5	the	O
6	wage	O
7	increase	O
8	,	O
9	that	PRICES
10	same	PRICES
11	basket	PRICES
12	cost	PRICES
13	$	PRICES
14	315	PRICES
15	.	O

Table 10: Example for CoNLL-formatted aspect data with preceding topic information

Parameter	**Value**
Learning rate	5.0e-6
Max epochs	50
Batch size	16
Scheduler	Linear with warmup
Warmup ratio	0.1
Number of repeats	5

Table 11: Hyperparameters for all experiments. The other parameters were Flairs default parameters.