SC|M
Studies in Communication and Media

FULL PAPER

Few-shot learning for automated content analysis: Efficient coding of arguments and claims in the debate on arms deliveries to Ukraine

Few-Shot-Lernen für automatisierte Inhaltsanalyse: Effizientes Codieren von Argumenten und Behauptungen in der Debatte um Waffenlieferungen an die Ukraine

Jonas Rieger, Kostiantyn Yanchenko, Mattes Ruckdeschel, Gerret von Nordheim, Katharina Kleinen-von Königslöw & Gregor Wiedemann

Studies in Communication and Media, 13. Jg., 1/2024, S. 72–100, DOI: 10.5771/2192-4007-2024-1-72

Jonas Rieger (Dr.), Technische Universität Dortmund, Lehrstuhl für Wirtschafts- und Sozialstatistik, Vogelpothsweg 78, 44227 Dortmund, Germany. Contact: rieger(at)statistik.tu-dortmund.de. ORCID: https://orcid.org/0000-0002-0007-4478

Kostiantyn Yanchenko (Dr.), Universität Hamburg, Fachgebiet Journalistik/Kommunikationswissenschaft, Von-Melle-Park 5, 20146 Hamburg. Contact: kostiantyn.yanchenko(at)uni-hamburg.de

Mattes Ruckdeschel (M.Sc.), Leibniz-Institute for Media Research | Hans-Bredow-Institut (HBI), Rothenbaumchaussee 36, 20148 Hamburg. Contact: m.ruckdeschel(at)leibniz-hbi.de

Gerret von Nordheim (Dr.), Universität Hamburg, Fachgebiet Journalistik/Kommunikationswissenschaft, Von-Melle-Park 5, 20146 Hamburg. Contact: gerret.vonnordheim(at)uni-hamburg.de. ORCID: http://orcid.org/0000-0001-7553-3838

Katharina Kleinen-von Königslöw (Prof. Dr.), Universität Hamburg, Fachgebiet Journalistik/Kommunikationswissenschaft, Von-Melle-Park 5, 20146 Hamburg. Contact: katharina.kleinen(at)uni-hamburg.de

Gregor Wiedemann (Dr.), Leibniz-Institute for Media Research | Hans-Bredow-Institut (HBI), Rothenbaumchaussee 36, 20148 Hamburg. Contact: g.wiedemann(at)leibniz-hbi.de. ORCID: https://orcid.org/0000-0002-4239-295X

FULL PAPER

Few-shot learning for automated content analysis: Efficient coding of arguments and claims in the debate on arms deliveries to Ukraine

Few-Shot-Lernen für automatisierte Inhaltsanalyse: Effizientes Codieren von Argumenten und Forderungen in der Debatte um Waffenlieferungen an die Ukraine

Jonas Rieger, Kostiantyn Yanchenko, Mattes Ruckdeschel, Gerret von Nordheim, Katharina Kleinen-von Königslöw & Gregor Wiedemann

Abstract: Pre-trained language models (PLM) based on transformer neural networks developed in the field of natural language processing (NLP) offer great opportunities to improve automatic content analysis in communication science, especially for the coding of complex semantic categories in large datasets via supervised machine learning. However, three characteristics so far impeded the widespread adoption of the methods in the applying disciplines: the dominance of English language models in NLP research, the necessary computing resources, and the effort required to produce training data to fine-tune PLMs. In this study, we address these challenges by using a multilingual transformer model in combination with the adapter extension to transformers, and few-shot learning methods. We test our approach on a realistic use case from communication science to automatically detect claims and arguments together with their stance in the German news debate on arms deliveries to Ukraine. In three experiments, we evaluate (1) data preprocessing strategies and model variants for this task, (2) the performance of different few-shot learning methods, and (3) how well the best setup performs on varying training set sizes in terms of validity, reliability, replicability and reproducibility of the results. We find that our proposed combination of transformer adapters with pattern exploiting training provides a parameter-efficient and easily shareable alternative to fully fine-tuning PLMs. It performs on par in terms of validity, while overall, provides better properties for application in communication studies. The results also show that pre-fine-tuning for a task on a near-domain dataset leads to substantial improvement, in particular in the few-shot setting. Further, the results indicate that it is useful to bias the dataset away from the viewpoints of specific prominent individuals.

Keywords: Pre-trained language models, transformer adapters, claim, argument and stance detection, automatic media content analysis, quality metrics.

Zusammenfassung: Pre-Trained Language Models (PLM) auf der Basis von Neuronalen Netzen in der Transformer-Architektur wurden im Bereich des Natural Language Processing (NLP) entwickelt und bieten große Möglichkeiten – insbesondere zur Codierung komplexer semantischer Kategorien in großen Datensätzen durch überwachtes maschinelles Lernen – zur Verbesserung der automatischen Inhaltsanalyse in den Kommunikationswissenschaften. Drei Faktoren verhinderten jedoch eine breite Nutzung der Methoden in den anwendenden Disziplinen: die Dominanz englischsprachiger Modelle in der NLP-Forschung, die erforderlichen Rechenressourcen und der hohe Bedarf an Trainingsdaten für das Fine-Tuning der PLMs. In der vorliegenden Studie gehen wir diese Herausforderungen an, indem wir ein mehrsprachiges Transformer-Modell in Kombination mit der Adapter-Erweiterung und Few-Shot-Lernverfahren verwenden. Wir testen unseren Ansatz an einem realistischen Anwendungsfall aus der Kommunikationswissenschaft zur automatischen Erkennung von Forderungen und Argumenten sowie deren Haltung in der deutschen Mediendebatte über Waffenlieferungen an die Ukraine. In drei Experimenten evaluieren wir (1) verschiedene Datenvorverarbeitungsstrategien und Modellvarianten für diese Aufgabe, (2) die Performanz verschiedener Few-Shot-Lernmethoden, und (3) wie gut das beste Setup bei unterschiedlichen Trainingsmengen in Bezug auf Validität, Reliabilität, Replizierbarkeit und Reproduzierbarkeit der Ergebnisse abschneidet. Wir können zeigen, dass die von uns vorgeschlagene Kombination von Transformer-Adaptern mit Pattern-Exploiting-Training eine parametereffiziente und leicht zu teilende Alternative zum vollständigen Fine-Tuning von PLMs darstellt. Sie ist in Bezug auf die Validität gleichwertig und bietet insgesamt bessere Eigenschaften für die Anwendung in den Kommunikationswissenschaften. Die Ergebnisse zeigen auch, dass das Pre-Fine-Tuning für eine interessierende Fragestellung auf einem inhaltlich verwandtem Datensatz zu einer erheblichen Steigerung der Validität führt, insbesondere im Falle weniger Trainingsdaten. Darüber hinaus zeigen die Ergebnisse, dass es sinnvoll ist, den verwendeten Datensatz hinsichtlich der Haltungen prominenter Personen zu bereinigen.

Schlagwörter: Pre-Trained Language Models, Transformer Adapter, Erkennung von Forderungen, Argumenten und Haltungen, automatisierte Medien-Inhaltsanalyse, Qualitätsmaße.

1. Introduction

The advent of pre-trained language models based on transformer neural networks such as BERT (Devlin et al., 2019) has revolutionized the field of natural language processing (NLP). This novel approach has the potential to overcome the limitations of previous computational text research by effectively capturing complex semantic concepts with respect to the sequentiality and contextuality of natural language (Wiedemann & Fedtke, 2021). The successful transfer of linguistic and semantic knowledge from self-supervised pretraining with a so-called masked language model task, i.e., a cloze test for the neural network to predict randomly masked words in a given sentence, to any target task in NLP has fueled the hope that powerful text classifiers can also be built for automatic content analysis with comparatively little training data. This task is researched in NLP under the term *few-shot learning* (FSL). So far, the implementation of these new methods in communication science has been somewhat limited. The lack of accessible and efficient ways to fine-tune large language models on the practical-technical side and the lack of best practice cases on the methodological side seem to hinder the adoption and development of new standards. Both problems are interrelated and can only be solved together.

This paper introduces a new implementation that we believe is particularly suitable for the widespread adoption of transfer learning in communication studies: a parameter-efficient fine-tuning (PEFT) based on adapters in combination with semi-supervised pattern-exploiting training (PET). The adapter approach (Pfeiffer et al., 2020) allows for the fine-tuning of transformer-based classifiers more efficiently in terms of training time and model size. The PET approach (Schick & Schütze, 2021) allows for the significant improvement of classification performance on small training datasets by using not only labeled texts but also the semantic knowledge from the labels themselves to learn from. We demonstrate the potential of our approach by applying it to a prototypical communication studies problem – the identification of claims and arguments in the news media debate.

Claims are a subject of interest for communication scholars because they serve as the fundamental conveyors of a perspective, or stance, and can therefore be utilized for a systematic examination of balance in media reporting (de Bruycker & Beyers, 2015). Arguments, in turn, are important because they reveal the underlying reasoning and evidence used to support a claim, which can aid in the evaluation of the validity and strength of a given perspective presented in a media debate (Meyers et al., 2000). Past studies on argument mining have indicated that identifying claims and arguments following a rigid formal structure such as in legal documents or student essays (Daxenberger et al., 2017; Hüning et al., 2022) is considerably easier compared to the more complex and latent argumentative structures found in news media debates. The goal of this paper is to show how traditional content analytic approaches to identify claims and arguments can be aided by state-of-the-art methods from supervised learning with a special focus on the promises and limitations of FSL. Along with this, we address the associated validity, reliability, reproducibility, and replicability concerns.

The proposed approach is demonstrated through a case study of the *Waffenlieferung* (arms delivery) debate in Germany. The issue of potential arms deliveries to Ukraine gained prominence in German news media several weeks before Russia's full-scale invasion of Ukraine and has since remained a recurring topic (Maurer et al., 2023). Over time, the *Waffenlieferung* issue proved to be both sensitive and complex and provoked various meta-debates, such as the alleged one-sided coverage of the issue by the German mainstream media (Precht & Welzer, 2022). At the same time, just as many other highly polarized issues such as vaccination or xenophobia, the *Waffenlieferung* debate implies one of the two opposing viewpoints – either to supply arms or not. From this perspective, the task of detecting claims and arguments on this topic can be considered basic compared to other more nuanced and multi-polar debates. Therefore, our case serves as an ideal starting point for exploring the capabilities of language models in identifying claims and arguments in news media. The straightforward polarity of the arms delivery debate allows us to focus on the core competencies of these models before future research may venture into more nuanced and intricate discourses.

At a broader level, our study also aims to overcome the three barriers to the more widespread adoption of computational methods in communication science identified by Baden et al. (2022): a) Using a real-life communication research use-case allows us to focus our method development from the outset on the requirements of measurement validity; b) moving beyond previous studies using automated analysis to classify either entire documents or identify specific expressions within texts, we capture arguments, claims, and stances as well as the relationships between them on a sentence level, allowing us to understand the argumentative complexity of a text; c) by working with multi-lingual transformer models the usefulness and applicability of our method developments are not limited to English language texts.

The rest of the contribution is structured as follows: Section 2 reviews the concepts of argument, claim, and FSL; Section 3 outlines our approach to automated identification of arguments and claims, including data, codebook, and language models; Section 4 presents our findings; and Section 5 discusses the implications of our study, including the key insights for communication scholars, limitations, and future research directions. Analysis scripts, models, and results as well as further technical details can be accessed at the corresponding OSF repository.[1]

2. Related work

The findings presented in this paper build upon a couple of research fields covering the conceptual definition of claim and argument, argument mining with the facet of stance detection, and FSL for text classification.

2.1 Claims and arguments

The concepts of claim and argument are prominent in NLP and machine learning fields where argument mining has been an important classification task since the beginning of the 2010s (Cabrio & Villata, 2018; Lippi & Torroni, 2016). At the same time, these concepts are less prevalent among communication scholars despite their relevance to many seminal topics in communication research. One such topic is the viewpoint diversity of the media discourse which is commonly approached from the perspective of article-level frames (Baden & Springer, 2017; Masini & Aelst, 2017) but can also be studied at the sentence level (Voakes et al., 1996), with each unique argument representing a different viewpoint or microframe. Another example would be the relevance of (unsupported) claims and (misleading) arguments for disinformation research (Vlachos & Riedel, 2014; Cook et al., 2017).

Conceptually, both claims and arguments originate from Argumentation Theory and can be operationalized differently depending on the research domain and school of thought one subscribes to (for a comprehensive overview of various theoretical approaches to the topic, see van Eemeren et al., 2014). As it is the goal in this study to automatically detect claims and arguments drawing on the common-sense perspective of a media text's reader, we will further focus on the basic claim-premise model that underlies most existing definitions of an argument (Goodman, 2018, p. 593) and has long been accepted in NLP (Walton, 2009).

[1] https://osf.io/tayxq

According to the claim-premise model, a *claim* – sometimes also referred to as *conclusion*, *standpoint*, *thesis*, or *proposition* – is an unjustified assertion about reality or, as Biran and Rambow put it, an "utterance which conveys subjective information" (2011, p. 364). When considered in the context of a polarized debate implying two opposing viewpoints, a claim can be classified by its stance as a claim *for* or *against* certain actions, policies, etc. Such a type of claim is called prescriptive (for the typology of claims, see Eemeren et al., 2014, p. 14). A claim is a self-sufficient unit that can be studied independently but is also a key component of an argument. To be considered an argument, a claim needs to be supported by one or several *premises*, sometimes also called *data*, *evidence*, or *reasons* (Lippi & Torroni, 2016). Premises are statements – or more generally pieces of information – that provide justification for an otherwise unsupported claim (Hüning et al., 2022). Thus, an argument can be viewed as a "set of statements, one of which [the claim] is meant to be supported by the other(s) [the premise(s)]" (Munson et al., 2004, p. 5). Often, arguments can be identified by so-called *argument markers* (Eckle-Kohler et al., 2015) – specific signal words and expressions associated with reasoning, e.g., *because*, *therefore*, *since*, etc. However, in natural language, premises can also connect with claims without argument markers resulting in less manifest argumentative structures. Table 1 provides illustrations of claims for/against, as well as arguments for/against in the context of our exemplary study case.

Table 1. The examples of claims and arguments regarding arms deliveries to Ukraine from the German media debate

Label	Example
claim for	*The strongest demands for arms deliveries come from the CDU.*
claim against	*Saxony's Prime Minister Michael Kretschmer (CDU) has spoken out vehemently against the delivery of heavy weapons to Ukraine.*
argument for	But if you really want to help the Ukrainians stop Putin's butchery troops and drive them out of the country, *you must now provide them with weapons* to go on the offensive.
argument against	This is due to the federal government not wanting to export armaments to crisis areas, and because of its past, *Berlin does not want to rush forward with arms deliveries* to the former Soviet Union.

Note. Claims are given in italics, premises are in regular font.

Two issues should be noted about the claim-premise model. First, the presented operationalization is deliberately minimalistic to match our practical interest in claims and arguments in news media discourse (for more extended versions of the claim-premise model, see Toulmin, 2003). Second, we acknowledge that argumentative structures are field- and context-dependent which means that the same utterances can take different functional roles depending on their position and meaning within the given text (Eemeren et al., 2014). To account for this, we incorporate contextual information when identifying claims and arguments both at the stages of coding (cf. Section 3.2) and model training (cf. Section 3.4).

2.2 Few-shot learning based on pre-trained language models

For the automated classification of claims and arguments given a set of labeled data, several large pre-trained language models (PLM) have been published in NLP such as RoBERTa (Liu et al., 2019) for English texts, as well as other PLMs specialized for different languages. For the processing of *arbitrary* languages or multilingual texts, the use of XLM-RoBERTa (Conneau et al., 2020) is well-suited. This offers a possibility to incorporate existing argumentative datasets in foreign languages as near-domain pre-training (Toledo-Ronen et al., 2020).

The common way of applying these models is standard fine-tuning on the corresponding downstream task (Devlin et al., 2019). By default, all parameters of the PLM are updated in a rather time- and resource-consuming process of gradient descent using the backpropagation algorithm. For this, fine-tuning requires a modern graphics processing unit (GPU) with sufficient memory. Later usage of the fine-tuned PLM is only possible by saving the entire model – for BERT and its successors a size of several gigabytes. Both characteristics may hinder the adaptation of the technology in applied science fields, where such hardware resources are often not readily available. To overcome these disadvantages, so-called adapter transformers have been proposed in NLP research (Pfeiffer et al., 2020). For adapters, model parameters of an initial PLM are frozen, and training is only performed on a set of additional parameter layers that are interwoven with the initial model. The first proposals for these additional layers were limited to bottleneck adapters, which mainly consist of two feed-forward neural nets with a down- and an up-projection. Several studies were able to confirm the authors' finding that this architecture performs on par with standard fine-tuning and mitigates negative effects such as the 'catastrophic forgetting' of pretrained knowledge (He et al., 2021).

For real-world applications such as automatic content analysis in communication science, adapters have two major advantages over standard fine-tuning. First, training time and model size can be reduced drastically because only the adapter module part of the entire model needs to be trained (cf. Table 7). Consequently, only the additionally trained parameters need to be stored to ensure model reproducibility. Due to their much smaller size, archiving, sharing, and reuse of trained models is greatly facilitated. Second, due to the regularization effect of freezing the initial PLM parameters, adapter training is less prone to overfitting than standard fine-tuning. This means that during training, for example, the number of epochs[2] can be chosen more liberally and does not require time-consuming tuning. It also means that performance results in repeated runs vary to a lesser extent due to random weight initializations than for standard fine-tuning.

Another common problem of real-world applications of supervised learning is the limited amount of training data. A special scenario for this, referred to as FSL in NLP, is pattern-exploiting training (PET; Schick & Schütze, 2021). The basic idea is that a model should not only learn from coded training data but also from

[2] The number of epochs refers to how often the learning algorithm iterates over the entire dataset in training.

instructions such as label names and label definitions to perform a classification task. In this scenario, a training example is injected together with distinct label patterns from the category codes into a so-called verbalizer statement, before being presented to the model. Instead of the prediction of pure (context-free) class labels (aka category labels for communication scientists), PET predicts masked tokens in the language modeling objective within predefined patterns (cf. Section 3.2). By this, it yields the pattern most likely filling a designated slot in the verbalizer that also contains the given training example. The authors were able to show that this approach also works well in realistic few-shot settings where prompt engineering and hyper-parameter tuning is not possible due to limited training data (Schick & Schütze, 2022). However, PET requires some human effort to define good pattern-verbalizer pairs.

3. Methodology

In our study, we demonstrate the potential of sequence classification using PLMs in the context of FSL for automated content analysis. For this, we develop a new approach to FSL combining two ideas from the NLP literature: adapters and PET. We compare our approach against other FSL methods (cf. Table 5) concerning their validity, reliability, replicability, reproducibility, and feasibility (cf. Table 2) for the task of predicting argumentative sentences (claim/argument) and their stance (for/against) on a dataset of newspaper articles on the topic of German arms deliveries to Ukraine.

Table 2. Definition of quality criteria in automated content analysis

Term	Description	Metric(s)
Validity	To what extent are the results consistent with a gold standard (here: human coder label)?	Accuracy, precision, recall, (macro-) F1
Reliability*	To what extent do repeated runs of the same methods on the same data produce similar results?	Deviation of validity metrics
Replicability*	To what extent does applying the same method to different (related) data produce similar results?	Deviation of validity metrics
Robustness*	To what extent do different parameter choices for the same model lead to different results?	Deviation of validity metrics
Reproducibility	Given the same model and data, (under what conditions) is it possible to obtain identical results?	Equality check, model size
Feasibility	Which resources are required for modeling?	Computation time, resource usage

Note. *We refer to these terms regarding validity assessment, while it is also possible to define them standalone on the predictions themselves. We do not investigate robustness in our experiments.

3.1 Data

The initial sample from the German media debate on arms deliveries to Ukraine included all articles mentioning *Ukraine* and *Waffen* [weapons] published between 1 January 2022 and 30 November 2022 in 145 German media outlets, a total of 26,057 articles (translation of German search terms in squared brackets). The corpus comes from the digital archive of the news magazine *Der Spiegel* and was originally compiled for a data journalism project. Within this initial corpus, we conducted additional searches using the combination of *Waffen* AND *L-/liefer* [deliver*] (in one sentence) OR militärisch* [military] & U-/unterstütz* [support] (in one sentence) and retrieved 14,697 articles for our closed corpus. To validate the search string (Mahl et al., 2022), 200 articles were randomly sampled from the initial corpus and labeled by two coders as either relevant or irrelevant to the topic of arms deliveries from Germany to Ukraine (Krippendorff's α = 0.902). Comparing the labels from manual coding with the labels generated by applying the search string revealed a precision score of 0.79 and a recall of 0.86, which were deemed adequate for our task. In a final step, the closed corpus was filtered further to include only leading German-language newspapers, both daily and weekly. This final corpus contained 7,301 articles from 22 media outlets (see Appendix 1 of the Supplementary Information file in the OSF) and was used to code claims and arguments.

3.2 Codebook

The codebook explicated how to systematically code arguments for/against and claims for/against arms deliveries to Ukraine, as defined on the previous pages. The four mutually exclusive codes were to be assigned at the sentence level. Both direct and reported claims/arguments were coded. Thus, sentences like "We must deliver arms now! – said X" and "X called for immediate arms supplies" were treated identically. Sentences that merely described the fact of arms deliveries and did not provide clues regarding the stance of an actor X on the topic were not coded, e.g.: "Germany can deliver 100 rocket-propelled grenades to Ukraine" (here, capability rather than a stance).

When coding for claims and arguments, coders were urged to always consider two contextual sentences before and after a potential claim/argument, which enabled them to correctly classify sentences that would have otherwise been left out. For instance, in a sequence of sentences "The flow of arms to Ukraine is enormous. This is unacceptable." the second sentence can only be coded as a claim against arms deliveries if contextual information is accounted for. Contextual information was also used later to train a classification model. Here, our communication science application scenario deviates from the standard argument mining task in NLP which usually operates on decontextualized sentences only (Jurkschat et al., 2022). This inclusion of context during manual coding needs to be reflected in modifications to procedures of machine input to the argument mining process (cf. Section 3.4).

Lastly, the codebook specified how to label those sentences that contained one actor contradicting another actor's position on arms deliveries, e.g.: "Plötner complained that the media were becoming more focused on tank deliveries instead of focusing on future relations with Moscow." The codebook referred to such sentences as *onion-structured* claims/arguments and instructed coders to always label the (alleged) stance of an actor on *arms deliveries* (here, media) rather than the stance of an actor on *another actor's position* (here, Plötner). To enable controlling for the stance of the onion-structured claims/arguments at the further stages, such sentences were labeled with a separate code (for the full codebook, see Appendix 2 of the Supplementary Information file in the OSF). We hypothesize that logical contradiction and critique of well-formed arguments might be difficult to grasp for the machine from a few samples only. In our experiments, we, thus, investigate whether the original labels (commenters' stance) or the labels with a swapped stance (actor's position commented on), while keeping the argument concept as labeled, are better suited for automated processing using FSL.

3.3 Coding process

The coding for arguments and claims was carried out by four coders in parallel. For comparable coding units, the texts were split into sentences using the NLTK tokenizer (Bird et al., 2009). After attending a training session where the coders familiarized themselves with the codebook and practiced assigning codes to data, they performed three rounds of coding on three sets of articles randomly sampled from the final corpus. Each coding round was followed by a group discussion during which inconsistently coded sentences were analyzed and coding rules clarified.

Table 3 reports the results of the inter-coder reliability tests for each of the coding rounds. Due to the complexity of the coding task, inter-coder reliability was assessed in two steps. In the first step, coders needed to agree on whether each specific sentence from a dataset contained a stance regarding weapons deliveries to Ukraine and hence was relevant for further classification (for the agreement scores, see Table 3). In the second step, the relevant sentences on which the majority of coders agreed in each round were classified as either argument for/against or claim for/against weapons deliveries. As can be seen from Table 3, in the first round of coding, the coders achieved "moderate" (Hughes, 2021, p. 417) agreement scores while in the last two rounds, the scores improved to >0.6, which is generally regarded as "significant" (Cabrio & Villata, 2018, p. 5428) or "substantial" (Hughes, 2021, p. 417) agreement.

Table 3. Results of the inter-coder reliability tests for each of the coding rounds

	n coders	*n* articles	*n* total sentences	*α* relevant sentences	*n* relevant sentences	*α* claim/ argument
Round I	4	50	5242	0.594	119	0.497
Round II	4	70	6729	0.634	81	0.667
Round III	4	70	5073	0.814	246	0.662

When using the labeled data to train a model, the majority rule was applied again to address the issue of imperfect inter-coder agreement. Particularly, only sentences on which most coders agreed at a second coding step qualified for the training dataset (this meant the agreement of two or three coders depending on how many coders labeled a sentence as *relevant* in the first step). This approach guarantees high-quality data for training purposes but also allows for some degree of disagreement among the coders so that not only *easy* cases of claims and arguments are picked to train and later assess the classifier.

Table 4. Number of observations per label and data set in dependence of stance-swapping for argumentative sentences labeled as onion-structured

Label	Original			Onion		
Set	train	dev	test	train	dev	test
argumentfor	46	4	13	49	4	14
argumentagainst	45	7	14	42	7	13
claimfor	101	13	24	101	13	23
claimagainst	81	14	19	81	14	20
nostance	1091	155	317	1091	155	317

For later modeling, it is beneficial if the imbalance of stance to nostance sentences is not too large. Therefore, we randomly sampled a set of nostance sentences from the total set so that they make up 80 percent of the total dataset. For the quality assessment of the experiments in Section 4, we divide the dataset into three parts: a train (70%), a dev (10%), and a test (20%) set. Table 4 provides the final number of sentences for each category depending on whether the labels are onion-swapped.

3.4 Language models

We compare different approaches to fine-tuning PLMs for text sequence classification in automated content analysis. The four underlying approaches are explained in Table 5 regarding their conceptual idea and in the following with respect to their methodology. Please see Appendix 3 of the Supplementary Information file in the OSF for detailed information and explanation regarding the concrete implementation of the models.

Table 5. Comparison of different approaches for fine-tuning PLMs

Approach	FSL	PEFT	Description
Full fine-tuning (FT)			Updating all parameters of the pre-trained language model (PLM) by learning relations between features and codes from a coded dataset.
Near-domain fine-tuning	X		Updating all parameters of the PLM by learning relations between features and codes from a coded dataset that has similarities to the dataset of interest in at least one component. Near-domain fine-tuning is performed as a preliminary step of actual fine-tuning on the dataset of interest.
Pattern-exploiting training (PET)	X		Updating all parameters of the PLM by learning relations between features and semantic representations of codes from a coded dataset using a language modeling objective instead of the standard classification objective.
Adapters		X	Freezing all parameters of the PLM, adding and updating only new parameters by learning relations between features and codes from a coded dataset.

Note. FSL = few-shot learning, PEFT = parameter-efficient fine-tuning.

All our experiments are based on the XLM-RoBERTa model in the large variant from Huggingface's transformer package (Wolf et al., 2020), which has a file size of 2.1 GB. A decisive advantage of the model is its multilingualism so that all presented analyses are easily transferable to other languages. To account for the context of sentences during coding, we construct the input to our standard models (without PET) as follows

Input := [target sentence] </s> [context before] </s> [context after]

where </s> represents a special separator token. The model has a maximum input sequence length of 512 tokens. Longer inputs are right-truncated to 512 tokens, so that the context after would be cut first, then context before, then the target sentence. The order ensures preservation of the most informative parts for overly long inputs. On our dataset, shortening affected only a few nostance target sentences. Placing the target sentence at the beginning also facilitates the model to particularly focus on the first tokens. Positioning between contexts, in contrast, would result in greater variance in the positioning of the tokens most relevant to the task.

As a baseline model, we use standard fine-tuning with a learning rate of 5e-6 together with a linear learning rate scheduler with warm-up. This setup makes fine-tuning less prone to overfitting. By empirical testing, we decided to run the training for 30 epochs.[3] As a direct comparison, we consider fine-tuning with identical parameters, using an already fine-tuned model on a near-domain dataset. For this, we use the well-known UKP-SAM dataset containing argumentative

[3] We tested the standard fine-tuning comparing the more common setup of 10 epochs vs. 30 epochs. Training for 10 epochs leads to significantly worse results (see Appendix 6 of the Supplementary Information file in the OSF).

English-language sentences from eight different topics together with a pro/contra/neutral stance label (Stab et al., 2018; Reimers et al., 2019) and train the XLM-RoBERTa model with a learning rate of 5e-6 for two epochs.

3.4.1 Pattern-exploiting training (PET)

Since fine-tuning requires a certain amount of data and a common problem in practical applications is the small amount of labeled data, we investigate to what extent the state-of-the-art FSL model PET (Schick & Schütze, 2021) adds value to the correct identification of argumentative sentences. PET requires an additional manual step for its application, namely the definition of pattern-verbalizer pairs (PVPs). We use two types of PVPs, one somewhat naive and one more elaborate, and train the model for 10 epochs using a learning rate of 1e-5.

3.4.2 Parameter-efficient fine-tuning with adapters

As a resource-efficient alternative to full fine-tuning, we also investigate the use of adapters in our experiments. There are many possibilities to use different adapter architectures. The most frequently investigated architecture is the *pfeiffer* adapter (Pfeiffer et al., 2020), which produces consistently promising results. So, we will concentrate on this architecture for our analysis. We train it with a reduction factor of 16 and a learning rate of 5e-5 over 30 epochs.[4] These values can be seen as a reasonable default that has reliably provided solid to very good results in previous studies. Hence, we use them for all bottleneck adapters presented in this paper.

3.4.3 Combining parameter-efficient fine-tuning and pattern exploiting-training

As a combination of the ideas of adapters and PET, we implement a new adapter with a PET-like classification head. For this, we only investigate the use of the naive PVP and combine a standard bottleneck adapter with our implementation of the PET-like classification head. Due to the modularity of adapters, it is also possible to combine this implementation with near-domain standard/adapter fine-tuning. For reasons of complexity, we only consider the combination of near-domain adapter fine-tuning and the PET-like head in our experiments.

4. Identifying argumentative sentences in the news media debate

We conducted three major experiments that analyze different facets and properties of model decisions to determine an optimal workflow for FSL, which are described in detail in the following three subsections.

In the first experiment, we investigated the influence of two preprocessing steps on the validity of the results. We compare the standard fine-tuning as a baseline with additional fine-tuning on the near-domain fine-tuned model. To shed light on

[4] Adapters are even less prone to overfitting than fully fine-tuned models for higher learning rates and higher numbers of epochs due to the freezing of the initial model.

how supervised learning can grasp complex argumentative structures, we compare the performance of both models with respect to whether sentences labeled with onion-structured contexts can be predicted more accurately with their actual or their swapped stances (for/against) (cf. Section 2.1). During our initial experiments, we observed systematic misclassifications related to the mention of certain person names. Thus, we also investigate this potential bias by randomly replacing all person names with other names before training. We compare the total of eight model variants in terms of overall macro-F1, precision, and recall, as well as the corresponding class-specific validity measures. For all models, we perform five repetitions each, so that we can also assess the reliability of the validity values via the models' uncertainty. The second experiment examines the application of all models and methods presented in Section 3.2 in a few-shot setting by evaluating different models at varying training set sizes. In addition to assessing the validity of the predictions, we again estimate their reliability in relation to specific models or numbers of training data by repeating them five times. In the third experiment, we test the most promising model from the second experiment with respect to its suitability in the real few-shot scenario.

These experiments provide insights into the validity, reliability, and replicability of the results of various state-of-the-art NLP models (including our own newly introduced combination of adapters and PET), as well as two preprocessing decisions strongly related to their application on the task of argument mining. In addition, in Section 4.4, we discuss the feasibility and model sizes in combination with the replicability of the applied models and their practical applicability in the field of communication science.

4.1 Biasing models away from person's stance

Initially, it may sound fallacious to try to prevent a model from learning that certain individuals are closely associated with a certain position. Indeed, for human interpretation the person with whom a statement appears matters, as well (Westerwick et al., 2017, p. 346). However, these individual positions may distort the model too much, so in the example of Annalena Baerbock (German Federal Minister of Foreign Affairs) – who is 90 percent of the cases connected to a positive stance on weapons delivery in our dataset – the person may predominate over the remaining elements of the sentence and only positive stance is predicted for those sentences. This poses several dangers with respect to the generalizability of the resulting model. If a person's position changes over time, but our model was only trained on data up to the time of the change, we might consistently misclassify statements from that person. Similarly, if our dataset only contains sentences with Annalena Baerbock as the actor, but the sentences to be predicted then also contain sentences with her as a reference, then the model will already be biased with respect to the classification of the sentence purely by the occurrence of this person.

For this reason, in the following, we consider the effect of a regularization preprocessing step that is intended to remove the information on the association of single individuals with positions. To do this, we use named entity recognition (NER) using the model *ner-german-large* from the flairNLP package (Akbik et al.,

2019) to identify individuals in our dataset and replace each occurrence of an individual with a random entry from the list of all occurring person entities. We also make sure that repeated occurrence of the same entity in one single sentence is replaced with the same (random) entity. In the following experiment, we distinguish between models with respect to the preprocessing step *Person* with the characteristics *original* and *shuffled*, where *original* means no changes of the initial data and *shuffled* refers to the application of the described procedure.

Another preprocessing step we consider is the stance reversal of sentences labeled as onion structured. The idea here is to investigate to what extent PLMs can recognize stance-reversing phrases (e.g., negations, distancing). In total, the data contains 49 sentences labeled as onion-structured (cf. Section 3.1). In the experiment, we distinguish between the original dataset (original) and the dataset with swapped labels (onion).

We consider the two presented models with full fine-tuning, i.e., standard fine-tuning (FT) and near-domain fine-tuning as a pre-step of full fine-tuning (FT SAM) as described in Section 3.2. Instead of a fixed number of epochs, we follow a common approach by training for 50 epochs on our labeled dataset and selecting the best epoch with respect to the macro-F1 score using the dev set. We use this best model for evaluation on the test set.

Table 6 shows the results of the eight combinations from the investigated models and the two preprocessing steps. We examine macro-F1 score and accuracy for general performance assessment, while additionally reporting class-dependent F1 score, precision, and recall.

Table 6. Comparison of performance measures for full fine-tuned models based on two preprocessing decisions

Person	original				shuffled			
Label	original		onion		original		onion	
Model	FT	FT SAM	FT	FT SAM	FT	FT SAM	FT	FT SAM
Overall: Macro-F1	0.594 ± .027	0.634 ± .036	0.608 ± .067	0.627 ± .039	0.585 ± .040	**0.660** **± .027**	0.566 ± .048	0.635 ± .026
Accuracy	0.906 ± .005	0.908 ± .008	0.902 ± .009	0.908 ± .009	0.901 ± .007	**0.916** **± .011**	0.893 ± .010	0.906 ± .011
argumentfor: F1	0.243 ± .077	0.339 ± .111	0.322 ± .170	0.356 ± .061	0.271 ± .138	**0.395** **± .085**	0.198 ± .166	0.391 ± .036
Precision	0.317 ± .020	0.366 ± .112	0.360 ± .227	0.412 ± .087	0.311 ± .159	0.460 ± .114	0.215 ± .176	**0.465** **± .084**
Recall	0.215 ± .102	0.323 ± .123	0.314 ± .173	0.329 ± .073	0.246 ± .132	**0.369** **± .132**	0.186 ± .160	0.343 ± .029
argumentagainst: F1	0.355 ± .063	0.478 ± .061	0.467 ± .087	0.479 ± .098	0.386 ± .064	**0.507** **± .069**	0.483 ± .147	0.501 ± .032
Precision	0.456 ± .097	0.576 ± .108	0.549 ± .065	0.613 ± .058	0.654 ± .195	**0.666** **± .094**	0.537 ± .188	0.629 ± .122
Recall	0.293 ± .053	0.413 ± .050	**0.443** **± .153**	0.400 ± .107	0.280 ± .050	0.413 ± .065	**0.443** **± .123**	0.429 ± .045
claimfor: F1	**0.689** **± .023**	0.649 ± .023	0.573 ± .057	0.588 ± .041	0.597 ± .060	0.670 ± .031	0.531 ± .020	0.604 ± .052
Precision	**0.632** **± .021**	0.574 ± .018	0.578 ± .080	0.537 ± .044	0.527 ± .083	0.607 ± .047	0.462 ± .017	0.570 ± .097
Recall	**0.758** **± .031**	0.750 ± .053	0.574 ± .051	0.652 ± .048	0.700 ± .049	0.750 ± .026	0.626 ± .044	0.661 ± .058
claimagainst: F1	0.699 ± .024	0.719 ± .035	0.695 ± .059	0.726 ± .021	0.685 ± .024	**0.744** **± .038**	0.637 ± .039	0.694 ± .045
Precision	0.602 ± .035	0.642 ± .029	0.603 ± .076	0.648 ± .036	0.584 ± .018	**0.676** **± .043**	0.607 ± .062	0.616 ± .035
Recall	**0.838** **± .049**	0.819 ± .047	0.836 ± .079	0.827 ± .034	0.829 ± .038	0.829 ± .038	0.673 ± .034	0.800 ± .084
nostance: F1	0.991 ± .005	0.997 ± .002	0.990 ± .003	0.992 ± .003	0.992 ± .004	0.990 ± .004	0.989 ± .005	0.989 ± .005
Precision	0.979 ± .006	0.973 ± .006	0.976 ± .004	0.980 ± .005	0.977 ± .004	0.981 ± .012	0.979 ± .006	0.977 ± .010
Recall	0.985 ± .001	0.985 ± .003	0.983 ± .001	0.986 ± .001	0.984 ± .002	0.985 ± .005	0.984 ± .002	0.983 ± .003

Note. Mean values ± standard deviation.

This shows that the model with the best overall performance is FT SAM without onion-swap but with person shuffling with a macro-F1 score of 66.0 percent and an accuracy of 91.6 percent. It should also be mentioned that accuracy is a somewhat problematic metric here, since due to the imbalance of labels the majority baseline to always predict nostance would already achieve an accuracy of 80.4 percent.

It turns out that the individual classes (categories) differ greatly in their performance. The label nostance is classified best (best F1: 0.997), while argumentfor performs worst (best F1: 0.395). This observation can be well described by the number of training data per label. Arguments are consistently predicted with higher precision than recall. For claims, the pattern is the other way around, with one exception. In an error analysis of a single sample run of the best-performing model, we found that this can be explained by the presence or absence of argument markers: while their presence leads to the high precision of arguments, their absence causes the models to assume a claim. Thus, the most frequent error in this sample run is that a claim is predicted where an argument is present (10/26 errors), while the stance is still correctly predicted. Moreover, among correctly classified arguments, 7 out of 11 have either argument markers or trigger words such as *argument* or *consensus* (for the comparison of correctly and incorrectly classified arguments, see Appendix 4 of the Supplementary Information file).

Overall, the results from the first experiment do not show clear findings regarding the question of whether it is better to use the original coded labels or the onion-swapped labels for the models. In contrast, person shuffling leads to better results, since the model seems to focus more on the actual argumentative part of a sentence than just the person stating it. Thus, for all further analyses, we will use the person shuffle preprocessing. In addition, the results show that pre-training with a near-domain dataset improves the performance of the models.

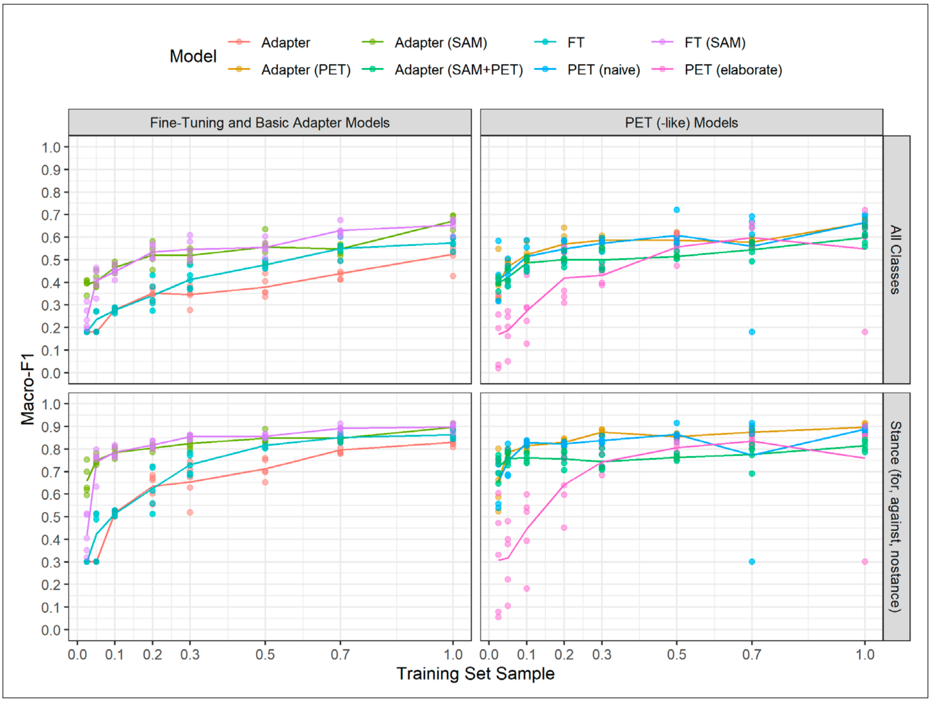
4.2 Model comparison for different training set sizes

Since in real applications, it is often not possible to label that much data (quickly) and since it is interesting to know what proportion of training data leads to sufficiently good predictions, we investigate eight different models with respect to their FSL performance regarding the overall macro-F1 score. For this purpose, we sample fixed train sets with proportions of 2.5, 5, 10, 20, 30, 50, 70, and 100 percent of the complete train set stratified by the labels (for absolute numbers of labels per train sample see Tables A2 and A3 in Appendix 5 of the Supplementary Information file in the OSF). In our FSL setting, we do not consider a dev set, but train with previously defined hyper-parameters as shown in Section 3.2, since such a dev set would not be available in a real FSL setting as well. We use person shuffle preprocessing and compare results from the dataset with the original labels.[5] All models are applied for five repetitions to assess the reliability of the models in the few-shot setting.

[5] We also ran the experiment for the onion-swapped labels. This yield comparable results, but no consistent improvement or decrease in performance.

In order to improve readability, we compare the four models based on general approaches in Figure 1 on the left and the four specialized FSL approaches on the right. In addition, from the original predictions on the 5-class problem in the top row, we also computed the performance for a pure stance prediction (for, against, nostance) in the bottom row.

Figure 1. FSL performance evaluation for different models and target labels



Note. Dots represent individual runs; curves represent mean values.

Confirming the result from the first experiment, the use of near-domain pre-training improves the performance of the models. In the FSL setting, we can see that for the adapter-based approaches, there is nearly a constant difference of 20 percentage points between the macro-F1 score for basic adapters and the additional use of a pre-trained adapter. The model *FT (SAM)* and *Adapter (SAM)* perform best among the general approaches. While the full fine-tuning on average provides slightly better results for the FSL settings with 30, 50, or 70 percent, the adapter-based approach achieves a macro-F1 score above 67 percent on the complete data set for both label types, which is better than the best model from Section 4.1, where also the dev set is used. The full fine-tuning after near-domain pre-training achieves similar scores as in the first experiment with a macro-F1 score of 0.653 (original) and 0.652 (onion). Considering the reliability of the validity estimation, both models can be considered equally well-suited for the task.

In the comparison of the standard and the pre-trained fine-tuning, there is not such a *constant* difference as for the adapter-based approaches, but the performances tend to become closer as the training data grows. Therefore, the pre-training with the SAM dataset is particularly beneficial in the FSL setting. Although we know that adapters perform on par on full data for several tasks (cf. Section 3.2), here the full fine-tuning approach performs slightly better than the basic adapter for growing training data.

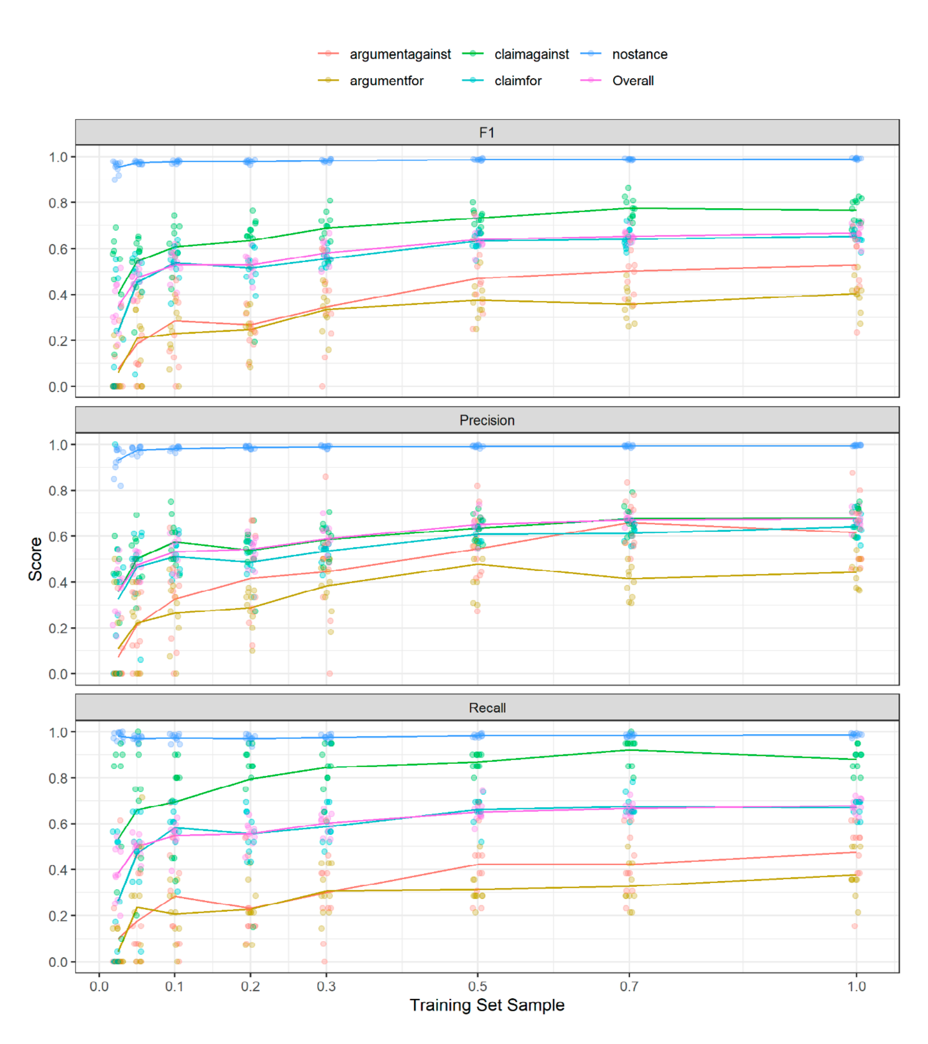
PET using the naive PVP, and our newly proposed PET head adapter perform overall on par, while on the full training data, our PET head performs best. The more elaborate PVP for a few training data is shown to perform poorly and at the same time to result in a high uncertainty in the validity of the results. Two outliers stand out for the naive PET for 70 percent and for the more elaborate PET for 100 percent of the training data. In both cases, the models end up predicting only nostance. For very few data, the combined *Adapter (SAM+PET)* model works quite well, but for a growing database, the performance of the model is almost consistently 10 percentage points below the *Adapter (PET)* model without SAM pre-training. Among the best-performing models (PET (naive), Adapter (PET), Adapter (SAM), and FT (SAM)) our newly proposed PET head has the lowest uncertainty in validity, i.e., the highest reliability of the results. Even for just five percent of the train set, our PET head achieves a macro-F1 score of almost 50 percent in the onion-swapped case. Finally, it can be seen that a combination of the SAM adapter and a PET head does not achieve the desired enhancing effect, but leads to a worsening of the results – (probably) due to the different training targets of the SAM adapter (classification) and the PET head (language modeling).

As already shown by the error analysis in the first experiment, the difficulty of the task consists in particular in the prediction of claim vs. argument, while the stance is estimated very reliably (macro-F1 of 90 percent on the complete data set). It should be noted that these models were not explicitly trained on the task of 3-class classification, and explicit training on this task could lead to (even) better results.

4.3 Argument mining in the true few-shot scenario

In the previous section, our adapter-based model with PET head proved to be the most promising in the FSL setting. In the following, we compare this model with respect to its validity and replicability in the true few-shot (TFS) setting. For this, we sample the train set not only once and stratified but ten times without the condition of stratification. By this, we simulate the process of coding random sentences without any prior knowledge of the actual label distribution. We examine the same sample proportions as in Section 4.2.

Figure 2. TFS performance evaluation of the PET head for differing subsets of the train set



Note. Dots represent individual runs; curves represent mean values. The dots are slightly jittered in their horizontal dimension for better readability.

Figure 2 shows how F1 score, precision, and recall relate to the amount of training data. The results show a difference in the overall macro-F1 score over ten repetitions of more than 30 percentage points between the worst and the best repetition in the case of only 2.5 percent of the training data. This uncertainty reduces to less than 20 percentage points for the FSL settings using 10, 20, 30, or 50 percent of the data, while for 70 percent it shrinks to nearly ten percentage points. As expected, the uncertainty decreases for higher proportions of the training set. It should also be noted that we cannot eliminate model uncertainty in this experiment. In Section 4.2, we found that even using the identical subsets for five repetitions in the case of 50 percent of the training data results in a range of five percentage points between the best and worst macro-F1 score for the PET head model we investigate here. This suggests that for a sufficiently large data set, the greater uncertainty comes from the modeling rather than from the data set sampling.

In addition, we can confirm the findings from Section 4.1 that, in principle, for arguments, the precision is higher than the corresponding recall, while for claims it is the other way around. The replicability of the results using the same method, but different samples of the same data basis is quite limited for this difficult task in the TFS setting. This is illustrated by the example of the reliability of the different validity measures, F1, precision, and recall, of which especially the latter two partly show large uncertainties.

4.4 Feasibility and reproducibility of the trained models

Besides validity (cf. macro-F1 scores in Sections 4.1, 4.2, and 4.3), reliability (cf. repetitions of models in Sections 4.1 and 4.2) and replicability (cf. different train sets in Section 4.3), feasibility and reproducibility are important properties for the (meaningful) applicability of automated models – not only for content analysis.

Table 7. Sizes and training time

Near-domain pre-training:	**Size**	**Runtime (100 %)**	**Runtime per epoch (100 %)**	**Runtime (2.5 %)**
SAM	2.1 GB	54 min	26.35 min	-
SAM Adapter	12 MB (+ 4 MB)	455 min	15.17 min	-
Fine-tuning:				
FT	2.1 GB	52 min	1.73 min	77 sec
Adapter	12 MB (+ 4 MB)	37 min	1.23 min	53 sec
Adapter (PET)	12 MB (+ 4 MB)	45 min	1.50 min	65 sec
PET	2.1 GB	54 min (+ 8 min)	5.40 min	77 sec (+ 8 min)

Note. The pre-training is done on the UKP-SAM dataset (cf. Section 3.2) while fine-tuning is performed using our presented dataset (cf. Section 3.1). The file sizes in parentheses refer to the classification head; the times in parentheses reflect the time PET needed to perform the language model objective in addition to the classification training. The experiments were conducted on a local machine using an NVIDIA GeForce RTX 3090 with 24GB.

In Table 7 the file sizes of the models are given, whereby sharing them guarantees the complete reproducibility of the results. Here it is assumed that the PLM XLM-RoBERTa is already available. All adapter approaches require only 12 MB each for storing all additional learned parameters of the model plus 4 MB for the prediction heads, whereas the parameters of the initial model remain the same. This makes them easy to share and just as easy to reproduce their predictions, especially in studies with numerous models. Combined with the finding that near-domain pre-training has a positive impact on the validity of the results, makes the (re)use of such near-domain adapters attractive.

In addition to the file size, the runtimes for training the models in the FSL scenarios are given for 2.5 and 100 percent of the train set. The near-domain pre-training requires considerable computation time due to the application to the rather large UKP-SAM dataset. However, we assume here that in practice one of these near-domain datasets and a pre-trained model are already available so the runtime in this case is not important for the actual application. The computation time for fine-tuning of our dataset, in contrast, is relevant and varies for the different models. At 5.40 minutes per epoch, PET is the most expensive, while the adapter-based approaches are significantly faster to compute at 1.23 minutes and 1.50 minutes per epoch, respectively. Thus, adapters form a viable option for application in low-resource environments. Compared to the adapters with standard classification heads, the PET head, at 45 minutes on the entire dataset, requires slightly more training time. However, its reliable and performant application without the need for tuning offers the possibility to obtain a decent model without running the training multiple times with different parameters.

5. Discussion

We investigate the use of PLMs for the identification of claims and arguments in (semi-) automated content analysis of media debates with little training data for a typical communication science research scenario. The application of three different experiments allows us to assess the validity of the results depending on preprocessing steps, model choice, and amount of training data. In addition, our experiments provide insights into the reliability, replicability, and reproducibility of the investigated model architectures, and allow us to draw some conclusions on the usefulness of this method for communication research.

5.1 General findings

In general, we found that all models studied are better at identifying manifest arguments and worse with latent argumentation, at least with the present amount of training data. The swapping of data labeled as onion-structured does not lead to a consistent difference in validity. A promising preprocessing step for the analysis of data in combination with stance is a random shuffling of individuals, which deprives the models of the information of the individuals' stances but puts more weight on other elements of the sentences. As a result, the validity of the results improved. Furthermore, it is shown that near-domain pre-training leads to better results than training on just in-domain data.

Despite their impressive performance, the models could not achieve satisfying values for all categories (aka classes) from a communication research (or machine learning) perspective. And yet, the validity – as well as their reliability over repetitions – should be sufficient to allow meaningful analyses of trends and distributions in the material based on the automatic classification alone.[6] In the category with the worst performance, arguments for weapons deliveries, most errors were a result of the models mistaking arguments for claims, with the stance almost always correctly predicted. In other words, even at this stage of training, the method explored here could already be used to rebuild all those previous studies which manually coded the distribution of pro- and contra-stances within media texts at a fraction of the cost. And even if at this point, manual coding may still be more reliable in identifying arguments for weapons deliveries, the time (and cognitive effort) required for this would be reduced substantially by first extracting the relevant sentences from the sample automatically using the proposed models.

5.2 Model choice

In the case of near-domain pre-training, adapters perform on par with full fine-tuning, while without pre-training, full fine-tuning becomes increasingly better than the selected basic adapter as the number of training data increases. Our newly proposed PET head performs on par with the original PET model in terms of validity but produces more reliable predictions. We were able to show that PET has a high uncertainty in the results when performed repeatedly on the same data, especially in the FSL setting and for too elaborated PVPs. Overall, for the presented task, our PET head is the best model choice based on validity, reliability, and reproducibility (parameter efficiency). For an application in low-resource environments, we recommend the use of adapters in combination with pre-training, for which also the manual step of engineering pattern-verbalizer pairs can be omitted.

5.3 Limitations

In our experiments, we have chosen selected and established default hyper-parameters. For adapters, we restricted our analysis to the best-known standard variant, the *pfeiffer* adapter in combination with a compression rate of 16. We did not investigate the robustness of the models with respect to the choice of parameters beyond the number of epochs in the FSL setting. We tested 10 epochs (a commonly chosen non-tuned parameter for full fine-tuning in NLP literature) against 30 epochs. Training with fewer epochs showed strong drops in the validity of the results (cf. Figure A1 in Appendix 6 of the Supplementary Information file in the OSF).

During the coding process, it became apparent how difficult the task is – even for human coders – so the question of possible error propagation compared to actual true labels (as a kind of *platinum* standard) arises. We have tried to counteract this by removing examples that cannot be clearly labeled to ensure the quality of the training data. However, it is unclear to what extent this decision

[6] Based on extensive experiments, Wiedemann (2019) concludes, “already moderate individual classifier performance regarding common evaluation measures such as F1 or Cohen’s kappa provide sufficiently accurate results to validly predict proportions and trends in a collection” (p. 155).

might lead to an overestimation of the validity of the results due to a potential simplification of the task.

5.4 Outlook

In further experiments, we will investigate the robustness of the PET head for different parameters and base architectures on different datasets. We aim to improve the implementation in such a way that PVPs might no longer be necessary (cf. von Werra et al., 2022; Mahabadi et al., 2022). Further potential for improvement may lie in the choice of the loss function used. Possibly the use of a triplet loss (cf. Sosnowski et al., 2022) in combination with certain data augmentation strategies could be promising.

Concerning the coding process, we plan to conduct several investigations. In terms of Active Learning (cf. Markus et al., 2023, Wiedemann 2019), it is interesting to find out whether there is potential for improvement during the coding process to draw further coding examples that are particularly useful for the model, using a more elaborate strategy than simple random sampling. We also aim to expand the coding of arguments beyond the sentence level by identifying premises among contextual sentences. This will not only allow us to provide a more complete picture of the German media debate concerning weapons deliveries to Ukraine but, more importantly, the inclusion of premises may help improve the models' performance in identifying arguments.

Moreover, since near-domain pre-training has proven to be useful and it has been shown in another work that stance prediction is not topic-independent (Reuver et al., 2021), it would be desirable to find the best fitting near-domain dataset, which is already thoroughly labeled, based on automated criteria and similarities to a weak-labeled dataset of interest. For this, we plan to develop a kind of similarity metric using different near-domain datasets that strongly correlate with their usefulness in being used as a near-domain dataset, so that it would be appropriate for finding a suitable dataset for pre-training.

Overall, by conducting these three experiments using the latest developments in NLP on a concrete communication science research case, we were able to address the three critical research gaps identified by Baden et al. (2022). Due to the close collaboration between computer and communication scientists within this project, it was possible to keep a very close eye on the validity of the classified constructs both by conducting manual error analyses but also by testing the impact of different preprocessing steps such as the de-biasing of the data by randomizing names. Furthermore, with the advent of multi-language transformer models such as the XLM-RoBERTa used here, we are no longer limited to the analysis of English texts (though of course languages without a sufficient amount of digital text available are still out of reach). And finally, we could show that by coding claims and arguments on a sentence level using the surrounding sentences as a context both for the human coders and the models, these methods can now be employed successfully for more complex constructs, greatly increasing the range of possible communication research questions that might be addressed.

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) and by the Bundesministerium für Bildung und Forschung (BMBF) as part of the project *FLACA: Few-shot learning for automated content analysis in communication science* (project no. 16DKWN064B).

References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 54–59. https://doi.org/10.18653/v1/N19-4010

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. C. G. (2022). Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures*, *16*(1), 1–18. https://doi.org/10.1080/19312458.2021.2015574

Baden, C., & Springer, N. (2017). Conceptualizing viewpoint diversity in news discourse. *Journalism*, *18*(2), 176–194. https://doi.org/10.1177/1464884915605028

Biran, O., & Rambow, O. (2011). Identifying justifications in written dialogs by classifying text as argumentative. *International Journal of Semantic Computing*, *5*(4), 363–381. https://doi.org/10.1142/S1793351X11001328

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit.* O'Reilly Media, Inc.

Cabrio, E., & Villata, S. (2018). Five years of argument mining: A data-driven analysis. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 5427–5433. https://doi.org/10.24963/ijcai.2018/766

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Cook, J., Lewandowsky, S., & Ecker, U. K. H. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLOS ONE*, *12*(5), e0175799. https://doi.org/10.1371/journal.pone.0175799

Daxenberger, J., Eger, S., Habernal, I., Stab, C., & Gurevych, I. (2017). What is the essence of a claim? Cross-domain claim identification. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2055–2066. https://doi.org/10.18653/v1/D17-1218

de Bruycker, I., & Beyers, J. (2015). Balanced or biased? Interest groups and legislative lobbying in the European news media. *Political Communication*, *32*(3), 453–474. https://doi.org/10.1080/10584609.2014.958259

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *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)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Eckle-Kohler, J., Kluge, R., & Gurevych, I. (2015). On the role of discourse markers for discriminating claims and premises in argumentative discourse. *Proceedings of the*

2015 Conference on Empirical Methods in Natural Language Processing, 2236–2242. https://doi.org/10.18653/v1/D15-1267

Eemeren, F. H. van, Garssen, B., Krabbe, E. C. W., Snoeck Henkemans, A. F., Verheij, B., & Wagemans, J. H. M. (2014). *Handbook of argumentation theory*. Springer Reference.

Goodman, J. (2018). On defining 'argument.' *Argumentation*, *32*(4), 589–602. https://doi.org/10.1007/s10503-018-9457-y

He, R., Liu, L., Ye, H., Tan, Q., Ding, B., Cheng, L., Low, J., Bing, L., & Si, L. (2021). On the effectiveness of adapter-based tuning for pretrained language model adaptation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2208–2222. https://doi.org/10.18653/v1/2021.acl-long.172

Hughes, J. (2021). krippendorffsalpha: An R package for measuring agreement using Krippendorff's alpha coefficient. *R Journal*, *13*(1), 413–425.

Hüning, H., Mechtenberg, L., & Wang, S. (2022). Detecting arguments and their positions in experimental communication data. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4052402

Jurkschat, L., Wiedemann, G., Heinrich, M., Ruckdeschel, M., & Torge, S. (2022). Few-shot learning for argument aspects of the nuclear energy debate. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 663–672. https://aclanthology.org/2022.lrec-1.69

Karimi Mahabadi, R., Zettlemoyer, L., Henderson, J., Mathias, L., Saeidi, M., Stoyanov, V., & Yazdani, M. (2022). Prompt-free and efficient few-shot learning with language models. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3638–3652. https://doi.org/10.18653/v1/2022.acl-long.254

Lippi, M., & Torroni, P. (2015). Context-independent claim detection for argument mining. *IJCAI'15: Proceedings of the 24th International Conference on Artificial Intelligence*, 185–191.

Lippi, M., & Torroni, P. (2016). Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology*, *16*(2), 1–25. https://doi.org/10.1145/2850417

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach* (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Mahl, D., von Nordheim, G., & Guenther, L. (2023). Noise pollution: A multi-step approach to assessing the consequences of (not) validating search terms on automated content analyses. *Digital Journalism*, *11*(2), 298–320. https://doi.org/10.1080/21670811.2022.2114920

Markus, D. K., Mor-Lan, G., Sheafer, T., & Shenhav, S. R. (2023). Leveraging researcher domain expertise to annotate concepts within imbalanced data. *Communication Methods and Measures*. Advance online publication. https://doi.org/10.1080/19312458.2023.2182278

Masini, A., & Van Aelst, P. (2017). Actor diversity and viewpoint diversity: Two of a kind? *Communications*, *42*(2). https://doi.org/10.1515/commun-2017-0017

Maurer, M., Haßler, J., & Jost, P. (2023). *Die Qualität der Medienberichterstattung über den Ukraine-Krieg* [The quality of media coverage of the Ukraine war]. Forschungsbericht für die Otto Brenner Stiftung. https://www.otto-brenner-stiftung.de/fileadmin/user_data/stiftung/02_Wissenschaftsportal/03_Publikationen/2023_Ukraine_Berichterstattung_Endbericht.pdf

Meyers, R. A., Brashers, D. E., & Hanner, J. (2000). Majority-minority influence: Identifying argumentative patterns and predicting argument-outcome links. *Journal of Communication*, *50*(4), 3–30. https://doi.org/10.1111/j.1460-2466.2000.tb02861.x

Munson, R., Conway, D., & Black, A. G. (2004). *The elements of reasoning* (4th ed). Wadsworth/Thomson.

Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., Cho, K., & Gurevych, I. (2020). AdapterHub: A framework for adapting transformers. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 46–54. https://doi.org/10.18653/v1/2020.emnlp-demos.7

Pfeiffer, J., Vulić, I., Gurevych, I., & Ruder, S. (2020). MAD-X: An adapter-based framework for multi-task cross-lingual transfer. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7654–7673. https://doi.org/10.18653/v1/2020.emnlp-main.617

Precht, R. D., & Welzer, H. (2022). *Die vierte Gewalt: Wie Mehrheitsmeinung gemacht wird, auch wenn sie keine ist* [The fourth estate: How majority opinion is made, even when it is not] (2. Auflage). S. Fischer.

Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., & Gurevych, I. (2019). Classification and clustering of arguments with contextualized word embeddings. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 567–578. https://doi.org/10.18653/v1/P19-1054

Reuver, M., Mattis, N., Sax, M., Verberne, S., Tintarev, N., Helberger, N., Moeller, J., Vrijenhoek, S., Fokkens, A., & van Atteveldt, W. (2021). Are we human, or are we users? The role of natural language processing in human-centric news recommenders that nudge users to diverse content. *Proceedings of the 1st Workshop on NLP for Positive Impact*, 47–59. https://doi.org/10.18653/v1/2021.nlp4posimpact-1.6

Schick, T., & Schütze, H. (2021). It's not just size that matters: Small language models are also few-shot learners. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2339–2352. https://doi.org/10.18653/v1/2021.naacl-main.185

Schick, T., & Schütze, H. (2022). True few-shot learning with prompts – A real-world perspective. *Transactions of the Association for Computational Linguistics*, *10*, 716–731. https://doi.org/10.1162/tacl_a_00485

Sosnowski, W., Wróblewska, A., & Gawrysiak, P. (2022). Applying SoftTriple loss for supervised language model fine tuning. *2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 141–147. https://doi.org/10.15439/2022F185

Stab, C., Miller, T., Schiller, B., Rai, P., & Gurevych, I. (2018). Cross-topic argument mining from heterogeneous sources. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3664–3674. https://doi.org/10.18653/v1/D18-1402

Toledo-Ronen, O., Orbach, M., Bilu, Y., Spector, A., & Slonim, N. (2020). Multilingual argument mining: Datasets and analysis. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 303–317. https://doi.org/10.18653/v1/2020.findings-emnlp.29

Toulmin, S. (2003). *The uses of argument* (Updated ed). Cambridge University Press.

Vlachos, A., & Riedel, S. (2014). Fact checking: Task definition and dataset construction. *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, 18–22. https://doi.org/10.3115/v1/W14-2508

Voakes, P. S., Kapfer, J., Kurpius, D., & Chern, D. S.-Y. (1996). Diversity in the news: A conceptual and methodological framework. *Journalism & Mass Communication Quarterly*, *73*(3), 582–593. https://doi.org/10.1177/107769909607300306

Von Werra, L., Tunstall, L., Thakur, A., Luccioni, S., Thrush, T., Piktus, A., Marty, F., Rajani, N., Mustar, V., & Ngo, H. (2022). Evaluate & evaluation on the hub: Better best practices for data and model measurements. *Proceedings of the 2022 Conference on*

Empirical Methods in Natural Language Processing: System Demonstrations, 128–136. https://aclanthology.org/2022.emnlp-demos.13

Walton, D. (2009). Argumentation theory: A very short introduction. In G. Simari & I. Rahwan (Eds.), *Argumentation in Artificial Intelligence* (pp. 1–22). Springer US. https://doi.org/10.1007/978-0-387-98197-0_1

Westerwick, A., Johnson, B. K., & Knobloch-Westerwick, S. (2017). Confirmation biases in selective exposure to political online information: Source bias vs. content bias. *Communication Monographs*, *84*(3), 343–364. https://doi.org/10.1080/03637751.2016.1272761

Wiedemann, G. (2019). Proportional classification revisited: Automatic content analysis of political manifestos using active learning. *Social Science Computer Review*, *37*(2), 135–159. https://doi.org/10.1177/0894439318758389

Wiedemann, G., & Fedtke, C. (2021). From frequency counts to contextualized word embeddings. In U. Engel, A. Quan-Haase, S. X. Liu, & L. Lyberg (Eds.), *Handbook of computational social science, Volume 2* (1st ed., pp. 366–385). Routledge. https://doi.org/10.4324/9781003025245-25

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., & Rush, A. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6