Limited Generalizability in Argument Mining: State-Of-The-Art Models Learn Datasets, Not Arguments

Marc Feger
Heinrich-Heine-University
Düsseldorf, Germany
marc.feger@hhu.de

Katarina Boland
Heinrich-Heine-University
Düsseldorf, Germany
katarina.boland@hhu.de

Stefan Dietze
GESIS - Leibniz Institute for the
Social Sciences & Heinrich-Heine-University
Düsseldorf, Germany
stefan.dietze@gesis.org

Abstract

Identifying arguments is a necessary prerequisite for various tasks in automated discourse analysis, particularly within contexts such as political debates, online discussions, and scientific reasoning. In addition to theoretical advances in understanding the constitution of arguments, a significant body of research has emerged around practical argument mining, supported by a growing number of publicly available datasets. On these benchmarks, BERT-like transformers have consistently performed best, reinforcing the belief that such models are broadly applicable across diverse contexts of debate. This study offers the first large-scale re-evaluation of such state-of-the-art models, with a specific focus on their ability to generalize in identifying arguments. We evaluate four transformers, three standard and one enhanced with contrastive pre-training for better generalization, on 17 English sentence-level datasets as most relevant to the task. Our findings show that, to varying degrees, these models tend to rely on lexical shortcuts tied to content words, suggesting that apparent progress may often be driven by dataset-specific cues rather than true task alignment. While the models achieve strong results on familiar benchmarks, their performance drops markedly when applied to unseen datasets. Nonetheless, incorporating both task-specific pre-training and joint benchmark training proves effective in enhancing both robustness and generalization.

1 Introduction

Undeniably, discourse gives people the opportunity to express and discuss their beliefs on any topic.

Argument mining, in this sense, is the automatic identification of the structure of inference and reasoning expressed as arguments presented in natural language (Lawrence and Reed, 2019).

Although there is no one-size-fits-all answer to *What is an argument?* (Stab et al., 2018), the idea suggests itself that arguments are latent yet observable and revolve around *how* they are constituted in terms of their logical scaffolding of argument discourse units, rather than *what* specific subject they address. In practice, these elements, whether sentences or sub-sentence segments, are pragmatically assigned functional roles, most commonly claims and premises, and form the fundamental building blocks of an argument (Stab and Gurevych, 2014; Daxenberger et al., 2017; Lawrence and Reed, 2019; Lopes Cardoso et al., 2023).

Consider the example *X should Y, because Z*, such as *Students should study, because it improves grades* or *We should reduce plastic use, because it minimizes ocean pollution*, which illustrates that the manifestation of an argument should ideally rely on structural components conveyed through functional patterns, while remaining agnostic of certain topics or other content-specific elements.

For this reason, one might assert that argument mining, in theory, is applicable across different corpora if the structural signals defining arguments are reliably identifiable from appropriately labeled data. Conversely, in practice, any inability to apply these signals to diverse datasets may expose systematic biases in the field, an issue that has long been informally discussed over coffee breaks.

Generalizability, in this regard, takes high priority, especially at leading NLP conferences such as ACL 2025, as it allows models to make reliable and reasonable predictions on data that does not correspond to their training data. This is especially true for real-world models, which should mimic human-like generalization abilities, where emerging evidence indicates that such models are often

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 23900–23915

fine-tuned to the specifics of established benchmark datasets, leading to unfounded optimism about their improvements (Saphra et al., 2024).

Consequently, concerns about vulnerability to shortcut learning (Geirhos et al., 2020) highlight the broader challenge of evaluating baselines beyond isolated benchmarks (Rendle et al., 2019).

Argument mining is one such area of natural language processing applications in which the ability to generalize is key. Hence, we ask for:

Q1: How comparable are the existing benchmark datasets for argument mining?

Q2: Do state-of-the-art argument mining models generalize to out-of-distribution data from other benchmarks?

Q3: Do these models acquire a generalizable concept of arguments?

In this context, there has been speculation that BERT (Devlin et al., 2019), known to pay great attention to basic syntax, nouns, and coreferences (Clark et al., 2019), is prone to learning shortcuts when mining arguments (Geirhos et al., 2020), where its generalization is limited to within-topic signals in datasets sharing similar argument and topic structures (Thorn Jakobsen et al., 2021).

Our aim is not to propose a new formalism for arguments or to pinpoint the best-performing argument mining model, but to use data from previous work in which different theories have been applied to see whether individual efforts and perspectives converge in terms of identifying arguments.

With this being said, we perform the first large-scale experimental assessment of benchmarks, systematically evaluating generalization across diverse argument mining datasets following a comprehensive review of datasets spanning 2008 to 2024.

For our study, we selected BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DistilBERT (Sanh et al., 2019) as exemplary BERT-like models, widely recognized as standard baselines in various areas of natural language processing (Rogers et al., 2020), including recent research on argument mining (Shnarch et al., 2020; Mayer et al., 2020a; Fromm et al., 2021a; Alhamzeh et al., 2022; Feger and Dietze, 2024b). We also examine WRAP (Feger and Dietze, 2024a), the only transformer whose language representation pre-training is extended by leveraging contrasts of inference and information signals to generalize argument components. Although originally designed for cross-topic generalization on Twitter (𝕏), WRAP does not rely on tweet- or topic-specific features to enhance its generalizability, distinguishing it from the others and making it particularly interesting for research.

In this study, we start by detailing our process of finding argument mining benchmark datasets and explain the selection criteria and justifications in Section 2. The core characteristics of these datasets, addressing research question **Q1**, are then examined in Section 3. Next, we describe our experimental setup in Section 4, covering both result generation and the implementation of best practices for significance testing, which form the basis for answering **Q2 - Q3** in Section 5. The results of this paper are then discussed in Section 6 and concluded in Section 7.

In order not only to elucidate the process but also to foster discussion that may inspire new approaches for novel datasets and broader generalization of argument mining methods, we contribute:

1. A survey of argument mining datasets between 2008 and 2024, primarily from the ACL Anthology, that identified 52 relevant papers with datasets from leading NLP conferences.

2. The first large-scale re-assessment that combines benchmark evaluations for 17 selected argument mining datasets, including controlled manipulation experiments to determine whether the reported state-of-the-art models (BERT, RoBERTa, DistilBERT, WRAP) actually learn generalizable argument concepts.

3. Statistical evidence that shortcut learning undermines generalization in argument mining. Although each of the examined transformers delivers strong results on benchmarks, all struggle to varying degrees when applied to other datasets, with WRAP generally performing slightly better. These challenges are compounded by divergent argument definitions and inconsistent annotations across datasets.

2 Argument Mining Benchmark Datasets

This section outlines the dataset collection and selection process, emphasizing the rationale behind our choice of benchmark datasets for argument mining. The decisions for all 52 datasets reviewed are present in Appendix A.1. Additionally, the code and data are available in our repository[1].

[1] Limited-Generalizability

Dataset	Paper	Genre	Definition	Arguments	No-Arguments
ACQUA	(Panchenko et al., 2019)	Mixed	Argumentative	1,949	5,236
WEBIS	(Al-Khatib et al., 2016a)	Online Debate	Argumentative	10,804	5,543
ABSTRCT	(Mayer et al., 2020b)	Academic	Claim-based	1,308	7,323
ARGUMINSCI	(Lauscher et al., 2018)	Academic	Claim-based	6,554	9,548
CE	(Rinott et al., 2015)	Encyclopedia	Claim-based	1,546	85,417
CMV	(Hidey et al., 2017)	Online Debate	Claim-based	979	1,593
FINARG	(Alhamzeh et al., 2022)	Spoken Debate	Claim-based	4,607	8,310
IAM	(Cheng et al., 2022)	Mixed	Claim-based	4,808	61,715
PE	(Stab and Gurevych, 2017)	Academic	Claim-based	2,093	4,958
SCIARK	(Fergadis et al., 2021)	Academic	Claim-based	1,191	10,503
USELEC	(Haddadan et al., 2019)	Spoken Debate	Claim-based	13,905	15,188
VACC	(Morante et al., 2020)	Online Debate	Claim-based	4,394	17,825
WTP	(Biran and Rambow, 2011)	Online Debate	Claim-based	1,135	7,274
AFS	(Misra et al., 2016)	Online Debate	Conclusion-based	5,150	1,036
UKP	(Stab et al., 2018)	Mixed	Evidence or Reasoning	11,126	13,978
AEC	(Swanson et al., 2015)	Online Debate	Implicit-Markup	4,001	1,374
TACO	(Feger and Dietze, 2024b)	Twitter Debate	Inference-Information	864	868

Table 1: The final 17 datasets that meet the sentential, binary label, and reproducibility criteria, each yielding at least 1,700 instances (850 per label) under a stratified 60/20/20 split, ensuring adequate size for the experiments.

2.1 Collection Process

As part of our data collection process, we examined the most recent and relevant survey papers on argument mining, primarily from the ACL Anthology (Daxenberger et al., 2017; Cabrio and Villata, 2018; Lawrence and Reed, 2019; Vecchi et al., 2021; Schaefer and Stede, 2021; Ajjour et al., 2023), all of which catalog datasets addressing various subtasks within the field, where argument identification is a fundamental prerequisite for each.

To expand and back up our dataset collection, we searched Google Scholar and Google Dataset Search for the keyword *argument mining* to find contributions beyond survey papers.

Based on our assessment, we found 52 such papers with datasets, mostly from top NLP conferences like ACL, NAACL, LREC, or EMNLP.

2.2 Selection Criteria

The dataset selection process for this paper was conducted in two stages. In the primary inclusion phase, we evaluated all 52 datasets based on:

- **Sentential**: The data and labels are at the sentence-level or aggregatable to this level (e.g., from sub-sentence or token annotations). Tweets were excluded from classical sentence conventions due to their unique structure.
- **Binary**: The dataset assigns binary labels to distinguish argument from no-argument sentences (e.g., based on the presence or absence of claims or other argument components).
- **Reproducible**: The dataset is largely replicable, with minor discrepancies from the publication (e.g., updates or duplicate removal affecting size). To ensure reproducibility, we reviewed documentation, labels, guidelines, and tools, and attempted to resolve access issues (e.g., client-sided or coding errors).

We applied these criteria sequentially, excluding datasets immediately upon failing any condition, eliminating 24 of the initial 52. In the refined inclusion step, we assessed relationships and data sufficiency to ensure adequate evaluation and generalization sizes, leading us to consider:

- **Related**: Connections between datasets such as updated versions, additional non-task-related features (e.g., stance added to a claim), and curated subsets derived from repositories that serve as data sources rather than datasets.
- **Sufficiency**: For a stratified 60/20/20 split, each dataset must have at least 500 training instances and 150 evaluation instances per label. An initial analysis revealed that two in five datasets fell short of this threshold, and alternative splits (e.g., 70/15/15 or 80/10/10) would further reduce evaluation sizes, worsening the small-data issue.

In total, this process resulted in 17 datasets encompassing ~345k labeled sentences, each meeting the aforementioned criteria. The final selection of datasets included in this study is listed in Table 1.

3 Characterizing Argument Mining Benchmark Datasets and Definitions

Before addressing **Q1**, we briefly introduce the individual datasets, organizing them by their primary labels. We then give the answer to **Q1** in terms of comparing definitions in Section 3.1 and textual characteristics in Section 3.2.

Argumentative serves as an umbrella term, identifying arguments with markers or patterns that suggest structural components, without necessarily specifying their roles (e.g., as claim or inference). In this sense, ACQUA (Panchenko et al., 2019) contains 7,185 argumentative sentences from Common Crawl (Panchenko et al., 2018), covering topics like computer science and brands, categorizing comparisons (e.g., Matlab vs. Python) as argumentative or not. Similarly, WEBIS (Al-Khatib et al., 2016a) comprises 16,347 segments across 14 topics (e.g., culture, health) from iDebate, with user-assigned labels (introduction, for, against) mapped to argumentative and non-argumentative labels.

Claim-based approaches explicitly annotate for the presence of claims as the core of an argument. Thereby, ABSTRCT (Mayer et al., 2020b), sourced from PubMed, comprises 8,631 sentences extracted from abstracts related to five diseases (e.g., neoplasm, glaucoma). ARGUMINSCI (Lauscher et al., 2018) provides annotations for the Dr. Inventor dataset (Fisas et al., 2016) for computer graphics publications, totaling 16,102 sentences. CE (Rinott et al., 2015) contains 86,963 sentences from Wikipedia across 58 topics (e.g., one-child policy, physical education). CMV (Hidey et al., 2017) consists of 2,572 sentences from the *Change My View* subreddit, spanning a diverse range of topics. FINARG (Alhamzeh et al., 2022) comprises 12,917 sentences sourced from transcribed earnings calls of Amazon, Apple, Microsoft, and Facebook. Moreover, IAM (Cheng et al., 2022) contains 66,523 sentences from various online platforms across 123 topics (e.g., vaccination, multiculturalism), while PE (Stab and Gurevych, 2017) includes 7,051 annotated sentences from persuasive essays (e.g., about cloning). SCIARK (Fergadis et al., 2021) contains 11,694 annotated sentences from scientific literature (e.g., PubMed, Semantic Scholar) on sustainable development goals (e.g., well-being, gender equality), also considering generalization to ABSTRCT. On the other hand, USELEC (Haddadan et al., 2019) offers 29,093 sentences from transcripts of U.S. presidential debates from 1960 (Kennedy vs. Nixon) to 2016 (Clinton vs. Trump), transcribed from the Commission on Presidential Debates. VACC (Morante et al., 2020) offers 22,219 sentences from a mixed collection of online debates about vaccination, while WTP (Biran and Rambow, 2011) includes 8,409 sentences from Wikipedia Talk Pages on various topics (e.g., Darwinism, the Catholic Church).

Others represents a residual category encompassing a variety of distinct definitions. AFS (Misra et al., 2016) comprises 6,186 annotated sentences drawn from online debate platforms such as iDebate and ProCon for three topics (e.g., gay marriage, death penalty). Sentences are labeled based on whether they explicitly convey a specific argument facet, with conclusions serving as the core component of the argument. UKP (Stab et al., 2018) contains 25,104 sentences across eight topics (e.g., nuclear energy, minimum wage) for cross-topic argument mining from heterogeneous sources, where arguments provide evidence or reasoning to support or oppose a topic. On the other hand, AEC (Swanson et al., 2015) contains 5,375 sentences on four topics (e.g., evolution, gun control) from CreateDebate, highlighting simple argument signals with labels based on the implicit markups: so, if, but, first, I agree that. Finally, TACO (Feger and Dietze, 2024b) comprises 1,734 tweets spanning six topics (e.g., abortion, Squid Game). It is designed for cross-topic argument mining on Twitter, focusing on inference to shape arguments.

3.1 Comparing Argument Definitions

(Q1) *Argument definitions vary, reflecting a spectrum of perspectives that contribute to a shared understanding of arguments.* Central to this is the observation that definitions mutually inform each other in their concepts (Lopes Cardoso et al., 2023). For example, in Table 1 most papers are claim-based, but when comparing the definitions, some view a claim as argumentative (Lauscher et al., 2018; Fergadis et al., 2021), others as conclusive (Mayer et al., 2020b), as stances (Rinott et al., 2015; Hidey et al., 2017; Cheng et al., 2022; Stab and Gurevych, 2017), or as a hybrid concept of all these (Haddadan et al., 2019; Morante et al., 2020).

Hence, further clarification is needed, especially concerning their generalization as part of **Q2 - Q3**. Thereby, Table 2, with examples from different definitions, illustrates whether their efforts nevertheless converge in the identification of arguments despite different perspectives.

Label	Dataset	Example
ARG	ACQUA	We chose MySQL over PostgreSQL primarily because it scales better and has embedded replication.
	SCIARK	In this case, if symptomatic, the treatment should be surgery, clinical follow-up, and counseling.
	AEC	So it would seem that if there is a scientific theory of [. . .], it has been tested [. . .] and therefore [. . .].
¬ARG	WEBIS	The Mo Ibrahim Prize was first established in 2007, and the prize represents [. . .] African leadership.
	FINARG	For those unable to attend in person, these events will be webcast and you can follow [...] at URL.
	TACO	'Bitter truth': EU chief [...] on idea of Brits keeping EU citizenship after #Brexit URL via USER

Table 2: Examples of argument (ARG) and no-argument (¬ARG) sentences from various datasets. Despite differences in definitions and topics, the similarities within and distinctions between label groups underscore the shared endeavor of argument mining approaches in identifying arguments, though each emerged differently.

3.2 Comparing Dataset Dimensions

First, the two text dimensions used to analyze the selected datasets are presented. For dataset-wise correlations of these, please refer to Appendix A.2.

Sentence-Level: To capture a broad, macro-level view without delving into individual word details, we used spaCy[2] to extract key textual attributes. These features reveal the overall structural and statistical properties of sentences, enabling sentence-level characterization of each dataset by:

- *Length*: Measured by the number of words per sentence, which serves as an indicator of linguistic complexity and verbosity.
- *Stop/Function Word Ratio*: The ratio of stop (e.g., it, is, are) and function words (e.g., against, because, therefore), including discourse markers, to the other words in a sentence to show their relative frequency of use.
- *Type-Token Ratio*: The ratio of unique words to total words in a sentence, assessing lexical diversity.
- *Readability*: The Flesch Reading Ease score quantifies text clarity, with lower values ($0 \leq$) indicating complex academic language and higher values (≤ 100) denoting easy readability, understandable by an 11-year-old.
- *Entropy*: Quantifies lexical unpredictability and the amount of information in a sentence, with values ranging from 0 (fully predictable text) to 1 (maximal unpredictability).
- *Sentiment*: Defined by polarity, ranging from -1 (extremely negative) to 1 (extremely positive), and subjectivity, ranging from 0 (objective) to 1 (subjective), possibly revealing persuasive strategies through emotions.
- *Part-of-Speech Tags*: The distribution of the 17 universal POS tags reflects basic syntax, lexical composition, and stylistic variation.

Word-Level: To compare datasets at the word level, we analyze the vocabulary of unique words used in each dataset. We extend this to words that convey the central semantic content of a sentence (e.g., government, abortion, freedom), that is, all words except stop and function words, discourse markers, and punctuation. Their relatedness or uniqueness is described using Jaccard similarity, a measure of similarity between two sets based on the ratio of their intersection to their union.

(Q1) *The sentence structures are strongly correlated across all datasets and labels.* On average, a sentence contains 21 words, with nearly every second word (48%) being a stop or function word. Sentences are lexically diverse (91% type-token ratio) yet highly readable (63% readability). The high predictability (22% entropy) and objective tone (43% subjectivity) suggest clear, structured writing with a slightly positive inclination (8% polarity). This is reinforced by the POS patterns, where sentences typically include five nouns, three punctuation marks, and two verbs, adpositions, and determiners, with other tags averaging below two.

Moreover, an average sentence closely aligns with both argument and no-argument sentences across these 24 sentence-level features (Spearman's $\rho \geq 0.97$), with a strong correlation ($\rho \geq 0.68$) across datasets. Slight differences exist in length, with an argument sentence averaging 24 words compared to 20 for a no-argument sentence, with readability scores of 60% and 64%, respectively.

(Q1) *Datasets and labels mainly differ in their semantic content.* Looking at the vocabularies, the datasets remain largely distinct, with 7–36% Jaccard similarity, a trend also observed for the semantic content words, reflecting their open-class.

[2] spacy.io

In contrast, stop, function, and discourse words show over 73% overlap due to their closed nature.

Interestingly, while comparing sentences across labels shows similar patterns, words describing the core semantic content remain largely distinct, overlapping below 48% and 19% on average, reinforcing lexical separation. Undeniably, the datasets share overlapping content, e.g., when discussing the one-child policy (PE) and abortion (IAM, TACO, UKP) or, figuratively speaking, the death penalty (AEC). Similarly, when discussing vaccination (VACC) overlaps might occur with medical (ABSTRCT) or sustainability (SCIARK) topics.

However, we found that these similarities are not very pronounced and that the datasets and labels are largely disjointed in terms of their core semantic content. This could provide the models with a shortcut opportunity, not based on how the labels are constructed, but rather on what they are about.

4 Experimental Setup

In this section, we outline the experimental setup and the best practices used for statistical testing to generate the data needed to answer **Q2 - Q3**.

Sampling: To create fixed training, development, and test sets, we used a 60/20/20 stratified split for each of the 17 datasets in Table 1, selecting 850 instances per label, corresponding to 1,700 samples per dataset and 28,900 in total.

Transformers: We selected BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DistilBERT (Sanh et al., 2019) as widely accepted standard baselines for NLP (Rogers et al., 2020), including argument mining (Shnarch et al., 2020; Mayer et al., 2020a; Fromm et al., 2021a; Alhamzeh et al., 2022; Feger and Dietze, 2024b). Further, we examined WRAP (Feger and Dietze, 2024a), the only transformer that is specifically pretrained for argument generalization. This applies contrastive learning to cluster similar manifestations of inference and information, separate dissimilar ones, and produce generalized embeddings robustly adaptable to downstream classification. However, our goal is to assess the generalizability of these state-of-the-art argument mining models, not to find the best. For these, we use the standard hyperparameter grid for GLUE (Wang et al., 2018), as accepted in the BERT and RoBERTa papers, balancing performance and time with a batch size of 32, 3 epochs, and a learning rate between 2e-5 and 5e-5, each trained on an A100 GPU.

Benchmarking and Generalization: The experiments presented here are the core investigations related to **Q2**. For each, we report the test results after tuning the hyperparameters to a target's development dataset, optimizing the macro F1 score to ensure equal importance of both labels.

We begin with an initial assessment using pairwise comparisons, following the transfer learning framework (Pan and Yang, 2010; Houlsby et al., 2019; Zhuang et al., 2019), where models are trained on one dataset and evaluated on others, including benchmarks on individual datasets. This yields a 17×17 matrix per model, with rows as training and columns as test data, see Figure 1.

Secondly, we conducted a supplementary experiment by training on all but one dataset and testing on the reserved one, forcing the models to generalize from joint benchmark data (Hays et al., 2023; Feger and Dietze, 2024a). Thereby, we will report the performance per model and evaluate each against the excluded dataset's state-of-the-art benchmark, compare Table 4 and Figure 1.

Disrupting Argument Signals: To build on the experiments addressing **Q2** and provide insight for **Q3**, we apply controlled input manipulation to both experiments described above. Specifically, we assess transformer performance after systematically removing stop and functional words (e.g., a, the, against, because), discourse markers, and punctuation using spaCy[2]. This process results in the elimination of around half the words in each sentence. It is therefore assumed that the removal of these lexical and syntactic elements, which also function as scaffolding for rhetorical and logical devices (Knott and Dale, 1994), suppresses the linguistic cues that, in theory, enable the distinction between the elements that constitute an argument and those that do not (Daxenberger et al., 2017; Opitz and Frank, 2019; Thorn Jakobsen et al., 2021). What remains is a lexical skeleton that primarily reflects topical and subject-related content while omitting functional and discursive elements, calling into question the model's ability to discern argued excerpts from mainly descriptive content (Lopes Cardoso et al., 2023), see Table 3.

Evaluation: We perform the experiments for **Q2 - Q3** and repeat them three times, each with varied samples and training initializations. To test significance, we use a two-way ANOVA with repeated measures for experimental robustness and one-tailed Student's t-tests for pairwise comparisons of models, see Appendix B for full details.

Label	Form	Example
ARG	Original	They should increase more routes to make people transport more easily.
	Manipulated	increase routes people transport easily
¬ARG	Original	Should governments spend more money on improving roads and highways?
	Manipulated	governments spend money improving roads highways

Table 3: Example from PE showing an argument (ARG) and no-argument (¬ARG) sentence in the original and manipulated form.

5 Results

In this section, we will address and answer questions **Q2 - Q3**. To this end, we will mainly focus on Figure 1, which compares the pairwise experiments to show which state-of-the-art argument mining model performs best, thus reflecting the current benchmark and generalization landscape. Tying in with this, we will then turn on Table 4 contrasting the state-of-the-art performance against those obtained by the models if trained on heterogeneous data. In addition, we elaborate on the insights gained from the controlled manipulations applied to these experiments. After that, we will discuss the significance of our results. However, for a better understanding, it can already be assumed that the results for each model and experiment follow a normal distribution, as confirmed with D'Agostino and Pearson's K^2 test ($p \geq .05$).

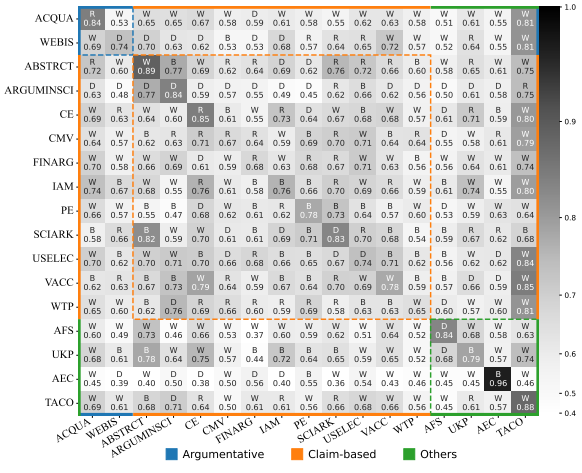

Figure 1: The best macro F1 scores from the benchmarking and pairwise generalization experiments, comparing WRAP (W), BERT (B), RoBERTa (R), and DistilBERT (D), indicate that strong performance is primarily achieved in the benchmark settings, as reflected along the main diagonal. Furthermore, WRAP excels in generalizing to TACO, as seen on the right.

(Q2) *Strong argument mining baselines do not necessarily imply strong argument generalization.* A notable observation in Figure 1 is the contrast between baselines on individual datasets and generalization across multiple datasets and definitions. Strikingly, 97% of generalization experiments fall below the mean benchmark result ($M = 0.79$), with 62% scoring under 0.65, while in 8% of cases generalization drops below 0.5 macro F1, highlighting the challenge of maintaining strong benchmark performances when tested on out-of-distribution datasets. We will further break down our answer:

Generalizability seems to be the exception rather than the norm. Given these circumstances, Table 1 shows several notable exceptions of good (≥ 0.75) to strong (≥ 0.8) generalizability across and within both definitional categories and genres, particularly for claim-based datasets. For instance, strong performance emerges within the academic domain, where SCIARK reaches 0.82 on ABSTRCT with BERT, and both ABSTRCT and ARGUMINSCI achieve 0.77 using BERT and DistilBERT. Evidence of cross-genre generalization also appears in cases such as IAM (mixed genre) and VACC (online debate), which achieve 0.76 and 0.79 on CE (encyclopedia) using RoBERTa and WRAP.

Broader generalization across definitions and genres is especially evident in UKP (evidence or reasoning, mixed), which surpasses 0.75 on both ABSTRCT (claim-based, academic) and CE (claim-based, encyclopedia) with BERT and WRAP. Similarly, TACO (inference-information, Twitter debate) consistently exceeds 0.8 across a vast range of definitions and genres with WRAP.

Still, both cross-definition and cross-genre generalization remain limited and exceptional.

Task-related pre-training appears to have a positive effect on overall performance and generalization. Numerically, WRAP ($M = 0.61, SD = 0.1$) shows the best overall performance in terms of macro F1. Notably, WRAP is the only model that attains a mean above 0.6 macro F1, while BERT ($M = 0.58, SD = 0.11$), RoBERTa ($M = 0.57, SD = 0.12$), and DistilBERT ($M = 0.56, SD = 0.11$) all perform worse. This performance advantage is particularly evident in cases where WRAP achieves the highest scores compared to the other models. In fact, WRAP demonstrates superior performance in 133 out of 289 experiments (46%), whereas BERT does so in 58 experiments (20%), RoBERTa in 50 experiments (17%), and DistilBERT in 48 experiments (17%).

	WRAP	BERT	RoBERTa	DistilBERT	SOTA	$\Delta_{max/min}$
ACQUA	**0.66**	0.6	*0.59*	*0.59*	**0.84**	0.18 / 0.25
WEBIS	0.63	**0.66**	*0.62*	0.65	**0.74**	0.08 / 0.12
ABSTRCT	**0.74**	**0.74**	**0.74**	*0.71*	**0.89**	0.15 / 0.18
ARGUMINSCI	**0.59**	*0.47*	0.55	0.5	**0.84**	0.25 / 0.37
CE	**0.77**	*0.72*	0.76	*0.72*	**0.85**	0.08 / 0.13
CMV	**0.63**	0.62	0.62	*0.58*	**0.67**	0.04 / 0.09
FINARG	*0.61*	0.62	**0.66**	0.65	**0.68**	0.02 / 0.07
IAM	**0.73**	*0.71*	**0.73**	**0.73**	**0.76**	0.03 / 0.05
PE	*0.65*	*0.65*	**0.69**	*0.65*	**0.78**	0.09 / 0.13
SCIARK	**0.75**	*0.73*	0.74	*0.73*	**0.83**	0.08 / 0.1
USELEC	**0.7**	0.66	0.68	*0.59*	**0.74**	0.04 / 0.15
VACC	*0.68*	**0.7**	*0.68*	0.69	**0.78**	0.08 / 0.1
WTP	**0.59**	0.55	0.55	*0.54*	**0.65**	0.06 / 0.11
AFS	*0.57*	0.58	0.59	**0.6**	**0.84**	0.24 / 0.27
UKP	**0.7**	*0.67*	**0.7**	0.68	**0.79**	0.09 / 0.12
AEC	0.52	**0.57**	*0.51*	0.56	**0.96**	0.39 / 0.45
TACO	**0.76**	0.61	0.65	*0.55*	**0.88**	0.12 / 0.33

Table 4: Transformers trained on all but the target benchmark are evaluated against their state-of-the-art baseline (SOTA), compare diagonal of Figure 1. *Minimum* and **Maximum** values indicate deviation from SOTA ($\Delta_{max/min}$). While all models fall short relative to SOTA, WRAP yields the best results in most cases.

Joint benchmark data for training may also help bootstrap reliable and improved generalization. Furthermore, the results of the supplementary experiment presented in Table 4 indicate that overall performance tends to improve when models are trained on joint benchmark data. Thereby, WRAP ($M = 0.66, SD = 0.07$), RoBERTa ($M = 0.65, SD = 0.07$), BERT ($M = 0.64, SD = 0.07$), and DistilBERT ($M = 0.63, SD = 0.07$) all achieve average macro F1 scores above 0.6, with values that are numerically higher than those observed in the pairwise setup. Again, WRAP shows the most consistent advantage, ranking first in 11 out of 17 experiments (65%).

(Q3) *State-of-the-art argument mining models are not solely defined by argument signals.* Following the controlled manipulation in the pairwise setup, all models dropped to similar levels, WRAP and BERT ($M = 0.56$, $SD = 0.09$), DistilBERT ($M = 0.55$, $SD = 0.1$), and RoBERTa ($M = 0.57$, $SD = 0.1$). Similar trends appear post-manipulation in the supplementary experiment for WRAP, RoBERTa, and DistilBERT ($M = 0.62$, $SD = 0.06$), and BERT ($M = 0.61$, $SD = 0.06$). With careful attention to detail:

Shortcut learning influences generalization of arguments, but task-related pre-training weakens the impact. For the pairwise experiments, BERT and DistilBERT showed almost no changes after manipulating inputs ($\Delta \leq 0.02$), while RoBERTa maintained its performance completely, suggesting that the overall performance of these models is not based on learning how arguments are constituted. In contrast, WRAP, which relies on its task-related pre-training to embed structural argument components across topics, showed the largest drop in macro F1 with $\Delta = 0.05$.

Jointly integrating benchmark data for training improves generalization and reduces shortcut reliance. The impact of WRAP towards robustness of generalization is also true for the supplementary experiment, where WRAP exhibited the largest performance drop ($\Delta = 0.04$) post-manipulation. Nonetheless, RoBERTa and BERT showed similar trends ($\Delta = 0.03$), while DistilBERT showed mostly no changes ($\Delta = 0.01$). Whereas the results in Table 4 show that each model underperformed relative to the state-of-the-art baselines, a notable pattern still emerged. This is, training on jointly integrated benchmark data raises the average macro F1 score to at least 0.64 for three out of four transformers and 0.63 for the lowest-performing model, compared to a maximum of 0.61 in pairwise transfer, achieved by WRAP. While only WRAP generalizes better in the pairwise setting and is less affected by lexical shortcuts, this advantage persists when trained on joined datasets. However, in this merged setting, RoBERTa and BERT also show improved robustness, despite their stronger reliance on shortcuts in the pairwise setup. Furthermore, average differences remain moderate with $\bar{\Delta}_{max} = 0.12$ and $\bar{\Delta}_{min} = 0.18$ while the models learn from heterogeneous data sources.

Differences in definitions of arguments reinforce the limitations of generalization. However, while signs of shortcut learning are found, it is undeniably not the sole limiting factor. Averaged across all models, misclassification patterns show that arguments are correctly classified 28% of the time and no-arguments 37%, suggesting that identifying no-arguments is easier. This is further supported by the lower misclassification rate for no-arguments (13%) compared to arguments (22%), highlighting practical differences in argument definitions that affect both generalization and benchmarks (e.g., due to conflicting annotations). This can also be observed when analyzing the misclassifications of individual models. Here, all models misclassify no-arguments as arguments in fewer than 16% of cases. In contrast, BERT, RoBERTa, and DistilBERT exhibit higher misclassification rates, ranging from 21% to 26%, while WRAP misclassifies arguments as no-arguments in 18% of cases, highlighting its superior generalization ability for arguments.

(Q2 - Q3) *The experiments demonstrate both statistical significance and practical relevance.* Repeated experiments support the robustness of these results. Regarding the pairwise experiments, a two-way repeated measures ANOVA for **Q2** showed a significant effect only when comparing model performances ($F(3, 864) = 69.47, \epsilon = 0.56, p_{\text{corr}} < .05, \eta^2_G = 0.03$), with negligible resampling or interaction effects. For **Q2**, paired one-tailed t-tests also showed that only model comparisons involving WRAP were significant ($p_{\text{corr}} < .05, 8.12 \leq t(288) \leq 10.14$), with moderate effect sizes ($0.39 \leq d \leq 0.49$). Similarly, repeating **Q3** revealed no significant effects, confirming that once ablated, the models perform comparably overall. Also, for **Q3**, when comparing pre- and post-manipulation results per model, only WRAP showed a relevant decrease ($p < .05, t(288) = -8.91, d = -0.49$). In terms of the supplementary experiments, repetition yielded no significant effects pre- and post-manipulation. However, regarding **Q3**, one-sided paired t-tests revealed significant post-manipulation decreases for WRAP, RoBERTa, and BERT ($p < .05, -5.52 \leq t(16) \leq -2.67, -0.58 \leq d \leq -0.41$), with WRAP showing the strongest effect.

6 Discussion

To summarize the limited generalization in argument mining addressed, Table 5 compares the best baseline results pre- and post-manipulation. On average, macro F1 differences remain close, within $\bar{\Delta}_{max} = 0.07$ and $\bar{\Delta}_{min} = 0.12$ per model, and in the best cases even exceed benchmark levels.

In the single case of AEC, which relies on only five keywords for arguments, overemphasis on these signals also appears to impair generalization. Although AEC attains the highest score (0.96) and experiences the largest post-manipulation drop (≤ 0.45, Table 5), its generalization is limited to 0.63 or even below 0.5, compare Figure 1. Given the low performance and minimal differences between pre- and post-manipulation results, BERT, RoBERTa, and DistilBERT do not clearly demonstrate an inherent ability to generalize arguments.

Although these challenges may be widespread, positive examples highlight the potential for future progress. This is particularly evident in cases involving diverse sources and topics (VACC, CE, TACO, UKP, IAM), where UKP, IAM, and TACO already aim for generalizable annotations.

	WRAP	BERT	RoBERTa	DistilBERT	SOTA	$\Delta_{max/min}$
ACQUA	*0.73*	0.77	0.76	**0.78**	**0.84**	0.06 / 0.11
WEBIS	*0.61*	0.66	0.66	**0.67**	**0.74**	0.07 / 0.13
ABSTRCT	*0.83*	**0.87**	0.84	**0.87**	**0.89**	0.02 / 0.06
ARGUMINSCI	0.78	**0.79**	*0.77*	*0.77*	**0.84**	0.05 / 0.07
CE	*0.75*	0.79	0.77	**0.81**	**0.85**	0.04 / 0.1
CMV	*0.57*	0.64	0.64	**0.65**	**0.67**	0.02 / 0.1
FINARG	0.62	*0.61*	0.66	**0.69**	**0.68**	-0.01 / 0.07
IAM	*0.66*	0.69	**0.71**	0.7	**0.76**	0.05 / 0.1
PE	*0.66*	0.67	0.71	**0.73**	**0.78**	0.05 / 0.12
SCIARK	*0.71*	**0.8**	0.77	0.79	**0.83**	0.03 / 0.12
USELEC	0.65	**0.66**	*0.62*	**0.66**	**0.74**	0.08 / 0.12
VACC	*0.67*	0.68	**0.69**	**0.69**	**0.78**	0.09 / 0.11
WTP	**0.58**	*0.54*	0.57	0.56	**0.65**	0.07 / 0.11
AFS	*0.78*	**0.81**	0.8	0.79	**0.84**	0.03 / 0.06
UKP	*0.74*	0.76	**0.78**	*0.74*	**0.79**	0.01 / 0.05
AEC	*0.51*	0.55	0.58	**0.59**	**0.96**	0.37 / 0.45
TACO	**0.77**	*0.76*	*0.76*	**0.77**	**0.88**	0.11 / 0.12

Table 5: Post-manipulation performance of each transformer compared to state-of-the-art (SOTA) results for baseline experiments per dataset. *Minimum* and **Maximum** values are highlighted, with $\Delta_{\text{max/min}}$ indicating their deviation from SOTA.

Despite limitations, the need for a unified structural approach to argument analysis becomes apparent. This is reinforced by the effectiveness of methodologies tailored to argument mining, as seen in WRAP's strong performance, averaging 0.75 when generalizing to TACO from all other datasets (Figure 1). Training on joint benchmark data further strengthens these abilities also for the standard transformers, even if numerical results fall short of the rarely doubted state-of-the-art (Table 4). Benchmarking should therefore build on combined datasets that capture the task's general demands, as in GLUE (Wang et al., 2018) and instruction-tuning benchmarks (Ouyang et al., 2022; Zhang et al., 2024), for which decoder-based argument mining (Cabessa et al., 2025) may be of interest.

7 Conclusion

We present the first large-scale re-evaluation of argument mining benchmarks through a generalization lens and evaluate whether the reported performance marks true progress. While structural patterns hold, thematic and content differences between labels and datasets favor shortcut learning. BERT, RoBERTa, and DistilBERT often rely on this to inflate benchmarks, while WRAP shows more resilience, likely due to its pre-training for argument generalization. Training on shared benchmark data further reduces shortcut reliance and improves generalization, notably in combination with WRAP. Our results stress the need to integrate different task demands and suggest re-framing argument mining as a joint generalizability task.

Limitations

This study did not separate direct from implicit arguments lacking clear structural and lexical cues, including discourse markers, and based on data analysis, assumed such cases are rare. However, this may affect interpretation, as implicit arguments are likely to depend on topical and content cues.

While we mostly used publicly available datasets, some require granted access.

Additionally, when extraction scripts were unavailable, we derived our procedures from both the available documentation and our understanding of the original process. This was particularly relevant for datasets where `.ann` files only provided annotated sequence boundaries for larger documents stored in `.txt` or `.json` formats. In such cases, we used spaCy[2] for sentence boundary extraction, which may produce boundaries that differ from the original assumptions. Nevertheless, we confirmed that over 95% of the extracted sentences ended with proper punctuation and began with a capital letter. We provide an extraction script[1] that automatically retrieves and processes all datasets considered.

The reproducibility of the experiments may be constrained by factors such as data size, runtime, and associated costs, with all experiments in this study running ~126 hours on a costly A100 GPU.

Acknowledgments

We sincerely thank the anonymous reviewers for their attentive and constructive feedback, which greatly contributed to improving the paper. Cheers!

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland. Association for Computational Linguistics.

Yamen Ajjour, Johannes Kiesel, Benno Stein, and Martin Potthast. 2023. Topic ontologies for arguments. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1411–1427, Dubrovnik, Croatia. Association for Computational Linguistics.

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data acquisition for argument search: The args.me corpus. In *KI 2019: Advances in Artificial Intelligence*, pages 48–59, Cham. Springer International Publishing.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016a. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1395–1404, San Diego, California. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016b. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443, Osaka, Japan. The COLING 2016 Organizing Committee.

Alaa Alhamzeh, Romain Fonck, Erwan Versmée, Elöd Egyed-Zsigmond, Harald Kosch, and Lionel Brunie. 2022. It's time to reason: Annotating argumentation structures in financial earnings calls: The FinArg dataset. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 163–169, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. Association for Computational Linguistics.

Or Biran and Owen Rambow. 2011. Identifying justifications in written dialogues by classifying text as argumentative. *International Journal of Semantic Computing*, 05(04):363–381.

Filip Boltužić and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, Baltimore, Maryland. Association for Computational Linguistics.

Jérémie Cabessa, Hugo Hernault, and Umer Mushtaq. 2025. Argument mining with fine-tuned large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6624–6635, Abu Dhabi, UAE. Association for Computational Linguistics.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 5427–5433. AAAI Press.

Liying Cheng, Lidong Bing, Ruidan He, Qian Yu, Yan Zhang, and Luo Si. 2022. IAM: A comprehensive

and large-scale dataset for integrated argument mining tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2277–2287, Dublin, Ireland. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Marc Feger and Stefan Dietze. 2024a. BERTweet's TACO fiesta: Contrasting flavors on the path of inference and information-driven argument mining on Twitter. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2256–2266, Mexico City, Mexico. Association for Computational Linguistics.

Marc Feger and Stefan Dietze. 2024b. TACO – Twitter arguments from COnversations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15522–15529, Torino, Italia. ELRA and ICCL.

Aris Fergadis, Dimitris Pappas, Antonia Karamolegkou, and Haris Papageorgiou. 2021. Argumentation mining in scientific literature for sustainable development. In *Proceedings of the 8th Workshop on Argument Mining*, pages 100–111, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Beatriz Fisas, Francesco Ronzano, and Horacio Saggion. 2016. A multi-layered annotated corpus of scientific papers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3081–3088, Portorož, Slovenia. European Language Resources Association (ELRA).

Michael Fromm, Evgeniy Faerman, Max Berrendorf, Siddharth Bhargava, Ruoxia Qi, Yao Zhang, Lukas Dennert, Sophia Selle, Yang Mao, and Thomas Seidl. 2021a. Argument mining driven analysis of peer-reviews. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6):4758–4766.

Michael Fromm, Evgeniy Faerman, Max Berrendorf, Siddharth Bhargava, Ruoxia Qi, Yao Zhang, Lukas Dennert, Sophia Selle, Yang Mao, and Thomas Seidl. 2021b. Argument mining driven analysis of peer-reviews. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6):4758–4766.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Nancy Green. 2018. Proposed method for annotation of scientific arguments in terms of semantic relations and argument schemes. In *Proceedings of the 5th Workshop on Argument Mining*, pages 105–110, Brussels, Belgium. Association for Computational Linguistics.

Giulia Grundler, Piera Santin, Andrea Galassi, Federico Galli, Francesco Godano, Francesca Lagioia, Elena Palmieri, Federico Ruggeri, Giovanni Sartor, and Paolo Torroni. 2022. Detecting arguments in CJEU decisions on fiscal state aid. In *Proceedings of the 9th Workshop on Argument Mining*, pages 143–157, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2023. Mining legal arguments in court decisions. *Artif. Intell. Law*, 32(3):1–38.

Ivan Habernal and Iryna Gurevych. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2127–2137, Lisbon, Portugal. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4684–4690, Florence, Italy. Association for Computational Linguistics.

Marcus Hansen and Daniel Hershcovich. 2022. A dataset of sustainable diet arguments on Twitter. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 40–58, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. 2022.

QT30: A corpus of argument and conflict in broadcast debate. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3291–3300, Marseille, France. European Language Resources Association.

Chris Hays, Zachary Schutzman, Manish Raghavan, Erin Walk, and Philipp Zimmer. 2023. Simplistic collection and labeling practices limit the utility of benchmark datasets for twitter bot detection. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 3660–3669, New York, NY, USA. Association for Computing Machinery.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Hospice Houngbo and Robert Mercer. 2014. An automated method to build a corpus of rhetorically-classified sentences in biomedical texts. In *Proceedings of the First Workshop on Argumentation Mining*, pages 19–23, Baltimore, Maryland. Association for Computational Linguistics.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. 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 2131–2137, Minneapolis, Minnesota. Association for Computational Linguistics.

Alistair Knott and Robert Dale. 1994. Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18(1):35–62.

Takahiro Kondo, Koki Washio, Katsuhiko Hayashi, and Yusuke Miyao. 2021. Bayesian argumentation-scheme networks: A probabilistic model of argument validity facilitated by argumentation schemes. In *Proceedings of the 8th Workshop on Argument Mining*, pages 112–124, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46, Brussels, Belgium. Association for Computational Linguistics.

John Lawrence, Floris Bex, Chris Reed, and Mark Snaith. 2012. Aifdb: Infrastructure for the argument web. In *Computational Models of Argument*, Frontiers in Artificial Intelligence and Applications.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018. Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2066–2081, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692.

Henrique Lopes Cardoso, Rui Sousa-Silva, Paula Carvalho, and Bruno Martins. 2023. Argumentation models and their use in corpus annotation: Practice, prospects, and challenges. *Natural Language Engineering*, 29(4):1150–1187.

Tobias Mayer, Elena Cabrio, Marco Lippi, Paolo Torroni, and Serena Villata. 2018. Argument mining on clinical trials. In *Computational Models of Argument*, Frontiers in Artificial Intelligence and Applications, pages 137–148.

Tobias Mayer, Elena Cabrio, and Serena Villata. 2020a. Transformer-based Argument Mining for Healthcare Applications. In *ECAI 2020 - 24th European Conference on Artificial Intelligence*, Santiago de Compostela / Online, Spain.

Tobias Mayer, Elena Cabrio, and Serena Villata. 2020b. Transformer-based argument mining for healthcare applications. In *European Conference on Artificial Intelligence*.

Rafael Mestre, Razvan Milicin, Stuart E. Middleton, Matt Ryan, Jiatong Zhu, and Timothy J. Norman. 2021. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *Proceedings of the 8th Workshop on Argument Mining*, pages 78–88, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Roser Morante, Chantal van Son, Isa Maks, and Piek Vossen. 2020. Annotating perspectives on vaccination. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4964–4973, Marseille, France. European Language Resources Association.

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument mining with structured SVMs and RNNs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 985–995, Vancouver, Canada. Association for Computational Linguistics.

Christopher Olshefski, Luca Lugini, Ravneet Singh, Diane Litman, and Amanda Godley. 2020. The discussion tracker corpus of collaborative argumentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1033–1043, Marseille, France. European Language Resources Association.

Juri Opitz and Anette Frank. 2019. Dissecting content and context in argumentative relation analysis. In *Proceedings of the 6th Workshop on Argument Mining*, pages 25–34, Florence, Italy. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359.

Alexander Panchenko, Alexander Bondarenko, Mirco Franzek, Matthias Hagen, and Chris Biemann. 2019. Categorizing comparative sentences. In *Proceedings of the 6th Workshop on Argument Mining*, pages 136–145, Florence, Italy. Association for Computational Linguistics.

Alexander Panchenko, Eugen Ruppert, Stefano Faralli, Simone P. Ponzetto, and Chris Biemann. 2018. Building a web-scale dependency-parsed corpus from CommonCrawl. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948, Lisbon, Portugal. Association for Computational Linguistics.

Prakash Poudyal, Jaromir Savelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. 2020. ECHR: Legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 67–75, Online. Association for Computational Linguistics.

Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie-Francine Moens. 2008. Language resources for studying argument. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics.

Steffen Rendle, Li Zhang, and Yehuda Koren. 2019. On the difficulty of evaluating baselines: A study on recommender systems. *ArXiv*, abs/1905.01395.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Naomi Saphra, Eve Fleisig, Kyunghyun Cho, and Adam Lopez. 2024. First tragedy, then parse: History repeats itself in the new era of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2310–2326, Mexico City, Mexico. Association for Computational Linguistics.

Robin Schaefer and Manfred Stede. 2021. Argument mining on twitter: A survey. *it - Information Technology*, 63(1):45–58.

Eyal Shnarch, Leshem Choshen, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2020. Unsupervised expressive rules provide explainability and assist human experts grasping new domains. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2678–2697, Online. Association for Computational Linguistics.

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

Ireland. Dublin City University and Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

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

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226, Prague, Czech Republic. Association for Computational Linguistics.

Milagro Teruel, Cristian Cardellino, Fernando Cardellino, Laura Alonso Alemany, and Serena Villata. 2018. Increasing argument annotation reproducibility by using inter-annotator agreement to improve guidelines. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

Terne Sasha Thorn Jakobsen, Maria Barrett, and Anders Søgaard. 2021. Spurious correlations in cross-topic argument mining. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 263–277, Online. Association for Computational Linguistics.

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

Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2020. Fine-grained argument unit recognition and classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9048–9056.

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. In *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)*, pages 1338–1352, Online. Association for Computational Linguistics.

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 812–817, Istanbul, Turkey. European Language Resources Association (ELRA).

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Michael Wojatzki and Torsten Zesch. 2016. Stance-based argument mining - modeling implicit argumentation using stance. In *Conference on Natural Language Processing*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024. Instruction tuning for large language models: A survey. *Preprint*, arXiv:2308.10792.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2019. A comprehensive survey on transfer learning. *CoRR*, abs/1911.02685.

A Extended Descriptive and Experimental Details

This appendix provides additional data and details omitted from Sections 2 and 3.

A.1 Section 2

For Section 2 we present the entire decision-making process for the selection of the benchmark datasets used in this work, which is in Table 6.

A.2 Section 3

Figure 2 extends the analysis in Section 3.2 by showing pairwise Spearman's ρ correlations for all reproducible datasets, including those omitted from experiments due to their small size.

Figure 3 extends the vocabulary analysis from Section 3.2 by displaying word overlaps across all datasets with available data.

B Statistical Design Protocol

In this appendix we also explain our protocol for the best-practices of statistical testing as described in Section 4 and applied in Section 5.

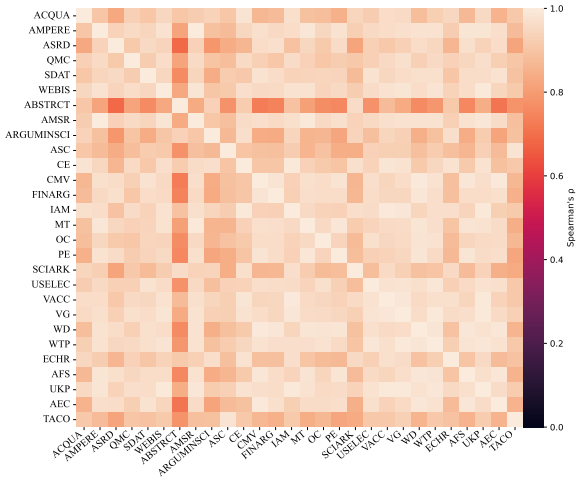

Figure 2: The correlations of the individual datasets (as well as the labels) in relation to the sentence-related features show a strong overall correlation ($\rho \geq 0.68$). Most strikingly, the ABSTRCT dataset stands out as medical texts exhibit different sentence structures from conventional ones, characterized by technical language, methodological details, and numerical values.

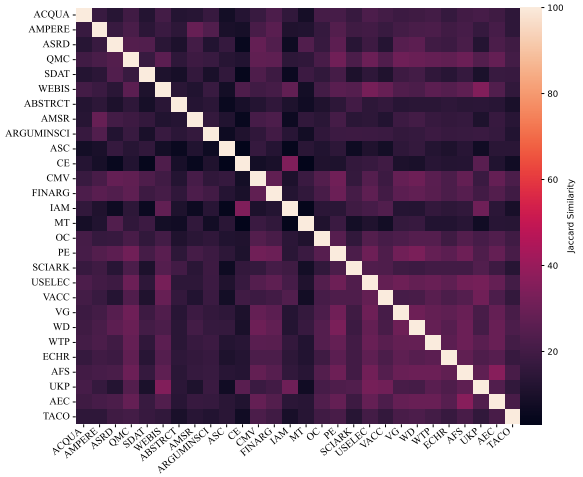

Figure 3: The word overlaps, measured by the Jaccard similarity between the vocabularies of two datasets, show that the datasets (as well as the labels) are generally distinct from each other. The overlaps range between 3–36%, with an average of 19%.

B.1 Two-Way Repeated Measures ANOVA

We employ a two-way repeated measures ANOVA to evaluate the effects of sampling (factor 1) and model choice (factor 2) on the macro F1 (dependent variable), with each dataset pair treated as a subject.

For valid inference, the following assumptions must be met:

- **Continuous Dependent Variable**: By definition, the macro F1 score is a continuous measure.
- **Within-Subject Design**: Each subject experiences every variation of both factors.
- **Normality**: The dependent variable is approximately normally distributed for each repeated measure (D'Agostino and Pearson's K^2 test).
- **Sphericity**: The variances of the differences between every pair of repeated measures are equal. If the Greenhouse-Geisser ϵ is below 0.75 (with values near 1 indicating compliance), we adjust the p-values (p_{corr}).

We can specifically evaluate for:

- **Sampling Effect**: Whether variations in data sampling (via different random seeds) influence model performance.
- **Model Choice Effect**: The performance differences among transformer models trained and evaluated on fixed samples. Each model is reinitialized in each trial using distinct random seeds to prevent carry-over effects.
- **Interaction Effect**: Whether the effect of sampling varies across the different models, offering insights into model stability under varying data conditions.

We evaluate the practical relevance of statistical significance using the effect size:

- **Generalized Eta Squared** (η^2_G): Proportion of the explained variance, interpreted as: ~0.01 (small), ~0.06 (moderate), ~0.14+ (strong).

B.2 One-Tailed Paired Student's t-Tests

Further, we conduct one-tailed paired t-tests as post-hoc analysis to identify directional differences (e.g., one model consistently outperforming another). These tests use the same assumptions as the prior ANOVA, except for sphericity. We apply the Bonferroni correction (p_{corr}) for multiple comparisons.

For these tests, we evaluate their practical relevance using the effect size:

- **Cohen's d**: The mean difference between paired conditions relative to the standard deviation of the differences, interpreted as: ~0.2 (small), ~0.5 (moderate), ~0.8+ (strong).

Dataset	Paper	Definition	Genre	Sent.	Binary	Reprod.	Related	Arg.	N-Arg.	Used
ACQUA	(Panchenko et al., 2019)	Argumentative	Mixed	Yes	Yes	Yes		1,949	5,236	Yes
AMPERE	(Hua et al., 2019)	Argumentative	Academic	Yes	Yes	Yes		6,729	242	No
ASRD	(Shnarch et al., 2020)	Argumentative	Spoken Debate	Yes	Yes	Yes		260	440	No
CDCP	(Niculae et al., 2017)	Argumentative	Online Debate	Yes	No					No
COMARG	(Boltužić and Šnajder, 2014)	Argumentative	Online Debate	No						No
EDIT	(Al-Khatib et al., 2016b)	Argumentative	Online Debate	Yes	No					No
IAC	(Walker et al., 2012)	Argumentative	Online Debate	No						No
MARG	(Mestre et al., 2021)	Argumentative	Spoken Debate	Yes	No					No
QMC	(Levy et al., 2018)	Argumentative	Encyclopedia	Yes	Yes	Yes		733	1,766	No
SDAT	(Hansen and Hershcovich, 2022)	Argumentative	Twitter Debate	Yes	Yes	Yes		387	210	No
WEBIS	(Al-Khatib et al., 2016a)	Argumentative	Online Debate	Yes	Yes	Yes		10,804	5,543	Yes
AAE	(Stab and Gurevych, 2014)	Claim-based	Academic	Yes	Yes	Yes	PE			No
ABSTRCT	(Mayer et al., 2020b)	Claim-based	Academic	Yes	Yes	Yes		1,308	7,323	Yes
AMECHR	(Teruel et al., 2018)	Claim-based	Legal	Yes	Yes	No				No
AMSR	(Fromm et al., 2021b)	Claim-based	Academic	Yes	Yes	Yes		839	561	No
ARGUMINSCI	(Lauscher et al., 2018)	Claim-based	Academic	Yes	Yes	Yes		6,554	9,548	Yes
ASC	(Wojatzki and Zesch, 2016)	Claim-based	Twitter Debate	Yes	Yes	Yes		147	568	No
CDC	(Aharoni et al., 2014)	Claim-based	Encyclopedia	Yes	Yes	Yes	CE			No
CE	(Rinott et al., 2015)	Claim-based	Encyclopedia	Yes	Yes	Yes		1,546	85,417	Yes
CMV	(Hidey et al., 2017)	Claim-based	Online Debate	Yes	Yes	Yes		979	1,593	Yes
CS	(Bar-Haim et al., 2017)	Claim-based	Encyclopedia	Yes	Yes	Yes	CE			No
DT	(Olshefski et al., 2020)	Claim-based	Spoken Debate	No						No
FINARG	(Alhamzeh et al., 2022)	Claim-based	Spoken Debate	Yes	Yes	Yes		4,607	8,310	Yes
IAM	(Cheng et al., 2022)	Claim-based	Mixed	Yes	Yes	Yes		4,808	61,715	Yes
MT	(Peldszus and Stede, 2015)	Claim-based	Microtext	Yes	Yes	Yes		112	337	No
OC	(Biran and Rambow, 2011)	Claim-based	Online Debate	Yes	Yes	Yes		702	7,824	No
PE	(Stab and Gurevych, 2017)	Claim-based	Academic	Yes	Yes	Yes		2,093	4,958	Yes
QT	(Hautli-Janisz et al., 2022)	Claim-based	Spoken Debate	Yes	No		AIFDB			No
RCT	(Mayer et al., 2018)	Claim-based	Academic	Yes	Yes	Yes	ABSTRCT			No
SCIARK	(Fergadis et al., 2021)	Claim-based	Academic	Yes	Yes	Yes		1,191	10,503	Yes
UGWD	(Habernal and Gurevych, 2017)	Claim-based	Online Debate	Yes	Yes	Yes	WD			No
USELEC	(Haddadan et al., 2019)	Claim-based	Spoken Debate	Yes	Yes	Yes		13,905	15,188	Yes
VACC	(Morante et al., 2020)	Claim-based	Online Debate	Yes	Yes	Yes		4,394	17,825	Yes
VG	(Reed et al., 2008)	Claim-based	Mixed	Yes	Yes	Yes	AIFDB	547	2,029	No
WD	(Habernal and Gurevych, 2015)	Claim-based	Online Debate	Yes	Yes	Yes		211	3,661	No
WTP	(Biran and Rambow, 2011)	Claim-based	Online Debate	Yes	Yes	Yes		1,135	7,274	Yes
ECHR	(Poudyal et al., 2020)	Conclusion-based	Legal	Yes	Yes	Yes		414	10,264	No
AFS	(Misra et al., 2016)	Conclusion-based	Online Debate	Yes	Yes	Yes	IAC	5,150	1,036	Yes
ARGSME	(Ajjour et al., 2019)	Conclusion-based	Online Debate	Yes	No					No
BASN	(Kondo et al., 2021)	Conclusion-based	Mixed	Yes	No					No
BIOARG	(Green, 2018)	Conclusion-based	Academic	Yes	No					No
DEMOSTHENES	(Grundler et al., 2022)	Conclusion-based	Legal	Yes	Yes	No				No
RSA	(Houngbo and Mercer, 2014)	Conclusion-based	Academic	Yes	No					No
AIFDB	(Lawrence et al., 2012)	AIF	Mixed	Yes	No					No
LAMECHR	(Habernal et al., 2023)	Custom Framework	Legal	Yes	No					No
ABAM	(Trautmann, 2020)	Evidence or Reasoning	Mixed	Yes	No		AURC			No
ASPECT	(Reimers et al., 2019)	Evidence or Reasoning	Mixed	Yes	No		UKP			No
AURC	(Trautmann et al., 2020)	Evidence or Reasoning	Mixed	Yes	Yes	No				No
BWS	(Thakur et al., 2021)	Evidence or Reasoning	Mixed	Yes	No		UKP			No
UKP	(Stab et al., 2018)	Evidence or Reasoning	Mixed	Yes	Yes	Yes		11,126	13,978	Yes
AEC	(Swanson et al., 2015)	Implicit-Markup	Online Debate	Yes	Yes	Yes	IAC	4,001	1,374	Yes
TACO	(Feger and Dietze, 2024b)	Inference-Information	Twitter Debate	Yes	Yes	Yes		864	868	Yes

Table 6: Summary of the 52 datasets from the reviewed papers, sorted by their applied definitions. Data collection followed the methodology described in Section 2.1, and selection criteria are detailed in Section 2.2. Empty entries indicate that the corresponding criteria were not further evaluated because a preceding criterion had already been rejected. The *Related* column indicates connections between datasets, like updates (e.g., AAE to PE, CDC to CE, RCT to ABSTRCT), additions of non-task-related features (e.g., CS adds stances to the claims from CE, ABAM adds aspects to the claims of AURC), or subsets from larger repositories (e.g., VG and QT from AIFDB, AEC and AFS from IAC).