Limited Generalizability: State-of-the-Art Models Exploit Content Signals, Bypassing Genuine Progress in Argument Mining

Anonymous ACL submission

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. Alongside theoretical insights into the structural components of arguments, a significant amount of research has focused on the practical extraction of arguments, leading to the growth of publicly available datasets where transformer models prevail and consistently attain highly competitive benchmark performance. Indeed, this has fostered the general assumption that argument identification is reliable and applicable in a variety of contexts. Our findings indicate that apparent progress, seen in near-ideal results, often arises from data limitations and labeling rather than inherent capabilities of the models. Specifically, we demonstrate that a shallow decision tree trained on a limited feature set, can generalize better then most state-of-the-art transformers. We emphasize the dependence of strong model performance and poor generalization on content-specific signals, which raises concerns about potential biases in existing research, issues of generalization, and the lack of standardization or simplification of argument extraction.

1 Introduction

Undeniably, a large part of societal discourse now occurs on the web, which has evolved into a vast public sphere where individuals express their beliefs and fulfill their desire to be heard across an ever-growing array of contexts and topics.

Argument mining, in this sense, refers to the automatic identification of arguments in natural language, serving as a crucial predecessor step to various related downstream tasks such as fact-checking, citation analysis, and opinion mining (Lawrence and Reed, 2019). Although there is no one-size-fits-all answer to the general question *What is an argument?* (Stab et al., 2018), we believe that arguments are latent, occurring in observable ways that depend on various factors with their manifestations reflecting a specific use of language. Our aim is not to propose a new formalism, but to use data from previous work in which different theories have been applied, and to combine these perspectives to see whether their individual annotations efforts converge in terms of identifying arguments.

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 definition of an argument, ideally, should rely on structural components conveyed through linguistic signals, while remaining agnostic of specific topics, domains, and other semantic elements.

Therefore, we argue that argument mining methods should also generalize well across corpora if they have learned to detect arguments from the labeled data, as failure to do so suggests biases.

For reference, early work addressed specific domains including legal texts (Moens et al., 2007), newspapers (Reed et al., 2008), essays (Stab and Gurevych, 2014; Persing and Ng, 2016; Wachsmuth et al., 2016), Wikipedia (Levy et al., 2014, 2017), user comments (Park and Cardie, 2014), dialogues (Swanson et al., 2015), and web discourses (Habernal and Gurevych, 2017).

As a result, there has been a growing interest in recent years in developing robust and generalizable argument mining techniques and datasets, particularly focused on the detection of arguments across topics from heterogeneous sources (Daxenberger et al., 2017; Levy et al., 2018; Stab et al., 2018).

Unsurprisingly, recent research has increasingly employed transformers for this challenge, with each claiming to achieve state-of-the-art results within different contexts, including transcribed debates (Shnarch et al., 2020), abstracts (Mayer et al., 2020), peer reviews (Fromm et al., 2021), financial calls (Alhamzeh et al., 2022), and diverse topics on Twitter (now 𝕏) (Feger and Dietze, 2024a,b).

This paper presents the first experimental assessment of argument mining methods and benchmarks, examining generalization across diverse datasets from 2015 to 2024. We investigate whether the reported state-of-the-art performance, especially from transformer models, signifies genuine progress, by assessing the extent to which they generalise across out-of-distribution data. We argue that models failing to generalize do not develop an inherent ability to distinguish arguments from non-arguments but are bound to spurious signals in the benchmark datasets. These may include superficial features like sequence length (Varis and Bojar, 2021), basic syntax, nouns, and co-references (Clark et al., 2019), which lead to an overemphasis on topic-specific words, such as 'Chernobyl' commonly appearing in discussions of 'nuclear energy', potentially hindering cross-topic generalization (Thorn Jakobsen et al., 2021).

Given that argument mining, by its very nature, initially demands the identification of arguments, through their structural components (Lawrence and Reed, 2019), models trained on such data should, in theory, demonstrate robust generalization across diverse datasets. However, it remains uncertain whether this theoretical generalization is consistently being achieved in practice, raising questions about the true effectiveness of models and data beyond isolated benchmarks (Rendle et al., 2019).

Following the transfer learning framework (Pan and Yang, 2010; Houlsby et al., 2019; Zhuang et al., 2019), we conduct pairwise comparisons by training on one dataset and evaluating on others, along with benchmark experiments on single datasets. We also assess robustness by training on all datasets except one and running an ablation experiment with minimal input context to determine if transformers selectively focus on content-specific signals or not.

These approaches include fine-tuning the state-of-the-art transformer models BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), DistilBERT (Sanh et al., 2019), and WRAPresentations (Feger and Dietze, 2024a), which is a specialized argument mining model designed for Twitter and usable for embedding argument components. Further, we challenge these transformers, which draw on linguistic patterns and word context, by incorporating basic features like length, readability, and others into a shallow decision tree, avoiding reliance on specific words and emphasizing the value of simpler features for better generalization across datasets (Joulin et al., 2017; Hays et al., 2023).

Thus, our investigation is guided by three aspects: *generalizability*, which explores the extent to which argument signals from specific source datasets can be applied across different targets; *simplicity*, which examines when simple features outperform transformers; and *dependencies*, to assess whether the datasets offer transformers primarily contextual word-label associations, introducing content biases, rather than argument signals.

Addressing these questions will shed light on the true progress in argument mining, may inform the creation of novel benchmark datasets and more robust argument mining models. To accomplish these objectives, we present the following contributions:

1. We combined 11 English benchmark datasets for sentence-level binary argument identification, forming the largest collection of its kind, covering diverse contexts relevant to argument mining, given an extensive literature review.

2. We are the first to conduct a comprehensive assessment of individual benchmarks and cross-dataset performances for argument identification, with ablation experiments to test robustness and analyze where transformers (BERT, RoBERTa, DistilBERT, WRAPresentations) focus their attention. All transformers from the original literature, reported as state-of-the-art, were challenged by a simple decision tree with basic linguistic features.

3. We observe that all models struggle to generalize across different datasets. However, the decision tree demonstrates slightly better and more consistent generalization, whereas transformers tend to focus on contextual word-label associations denoting a datasets content instead of capturing the structural components necessary for robust argument mining.

2 Argument Mining Benchmark Datasets

This section describes the dataset acquisition and selection process, explaining how the datasets were categorized based on the explicitly stated definitions for argument identification, and motivating our aim to investigate generalization across these. Additionally, the individual datasets and key textual features used for numeric representation are detailed. All datasets used in this study are listed in Table 1, with materials and reviews, including candidates, available in our public repository[1].

[1] anonymous.4open.science/r/Limited-Generalizability

Dataset	Reference	Definition	Context	ARG	¬ARG
acqua-2019	(Panchenko et al., 2019)	Argumentative	Common Crawl	1,949	5,236
ampere-2019	(Hua et al., 2019)	Argumentative	Peer Reviews	10,047	339
asrd-2020	(Shnarch et al., 2020)	Argumentative	Debate Transcripts	260	440
qmc-2018	(Levy et al., 2018)	Argumentative	Wikipedia	733	1,767
webis-2016	(Al-Khatib et al., 2016)	Argumentative	Debate Forums	10,846	5,556
abstrct-2020	(Mayer et al., 2020)	Claim-Premise	PubMed Abstracts	1,253	496
amsr-2021	(Fromm et al., 2021)	Claim-Premise	Peer Reviews	839	562
finarg-2022	(Alhamzeh et al., 2022)	Claim-Premise	Financial Calls	4,607	4,631
aec-2015	(Swanson et al., 2015)	Implicit Markup	Debate Forums	4,001	1,374
taco-2023	(Feger and Dietze, 2024b)	Inference-Information	Twitter Debates	865	869
ukp-2018	(Stab et al., 2018)	Evidence or Reasoning	Mixed Sources	11,126	13,978

Table 1: Collection of the 11 publicly available English sentence-level datasets used, organized by the *Argumentative*, *Claim-Premise*, and *Custom* definitions mentioned for argument identification. The table includes details on contexts and the distribution of arguments (ARG) and non-arguments (¬ARG). More details can be found in the repository[1].

2.1 Data Acquisition

To gather candidate datasets for our study, we used Google Dataset Search and Google Scholar, focusing on publicly available datasets with their related papers mentioning the term 'argument mining'.

We identified 69 search results, which were screened to ensure they included both a publication and an associated dataset containing English text, typically composed of single sentences. This process narrowed the selection to 26 candidate datasets, all published by top-tier NLP conferences (e.g., NAACL, ACL, EMNLP, LREC, COLING).

These 26 datasets were reviewed for binary labels distinguishing arguments (ARG) from non-arguments (¬ARG), ensuring the labels were provided by the datasets. We also assessed the availability of clear mapping rules for aggregating any sub-classes into these binary labels, alongside a documentation of data and the annotation process.

Datasets focused exclusively on pro/con classification (Walker et al., 2012), those limited to arguments but employing BIO tagging (Habernal et al., 2024), relation extraction (Mestre et al., 2021) or similarity tasks (Misra et al., 2016; Reimers et al., 2019; Thakur et al., 2021), as well as those that did not fit the quality criteria (Rinott et al., 2015; Wang, 2017; Habernal and Gurevych, 2017; Shnarch et al., 2018; Ein-Dor et al., 2019; Trautmann et al., 2020), those very small (Hansen and Hershcovich, 2022) or overly large (Cheng et al., 2022), were excluded.

This process yielded our final collection of 11 datasets that are 1) publicly available, 2) sentence-level English text, 3) spanning different definitions and contexts, and 4) containing binary labels for arguments (ARG) and non-arguments (¬ARG).

2.2 Combining Argument Definitions

In our data collection, we identified three categories from the explicitly stated definitions, applied by humans or heuristics. *Argumentative* serves as an umbrella term, identifying arguments with markers or patterns that suggest structural components, without necessarily specifying their roles. In contrast, the *Claim-Premise* approach explicitly annotates arguments, labeling claims as the central and premises as supporting elements for these. Finally, *Custom* describes those mentioning none of these.

Although definitions vary, arguments require a minimal component, often called a claim, reasoning, or inference, besides supporting elements like premises or evidence (Lawrence and Reed, 2019; Feger and Dietze, 2024b). Notably, qmc-2020 labels only *Claims*, omitting *Premises* but refers to them as *Argumentative*. aec-2015 uses *Argumentative* markers (e.g., 'it', 'so') but calls this *Implicit Markup*, while ukp-2018 and taco-2020 use *Evidence or Reasoning* and *Inference-Information*, resembling the *Claim-Premise* definition, suggesting exchange and overlaps with distinct variations.

Inconsistencies and differences in definitions are problems on their own and can affect generalization. Going beyond this, our focus is to assess whether content-specific biases in datasets hinder argument identification across different specific definitions, as their common purpose is to provide reliable ground truth for training models on this task and not to capture unintended correlations.

In combination with these points, we believe that efforts largely converge in labeling and distinguishing arguments and non-arguments, as shown by examples from different datasets, refer to Table 2.

Label	Dataset	Example
ARG	acqua-2019	We chose MySQL over PostgreSQL primarily because it scales better and has embedded replication.
	webis-2016	Clearly the presence of these weapons show how much worse the situation could get if assad is not overthrown.
	aec-2015	If guns are allowed at schools it should be the schools duty to teach the kids about the dos and donts with guns.
¬ARG	ampere-2019	If you want image-based pick and place with demonstrations for example, see Chebotar '17 (not cited).
	finarg-2022	For those unable to attend in person, these events will be webcast and you can follow the comments at microsoft.com/investor.
	taco-2023	'Bitter truth': EU chief pours cold water on idea of Brits keeping EU citizenship after #Brexit https://t.co/j3DteyWcMg via @TheLocalEurope

Table 2: Examples of arguments (ARG) and non-arguments (¬ARG) from different datasets, each representing a distinct definition (*Argumentative*, *Claim-Premise*, *Custom*), are presented in ascending order. Despite variations in definitions and intended purposes across datasets, the observed similarities and differences among label groups suggest a potential convergence in their applicability for argument identification, which, in theory, should promote generalization across datasets.

2.3 Dataset Collection

Argumentative. acqua-2019 (Panchenko et al., 2019), also referred to as compsent-19, consists of 7,185 argumentative sentences extracted from the Common Crawl (Panchenko et al., 2018) corpus. The dataset includes topics related to computer science, brands, and various other subjects. It focuses on classifying comparisons (e.g., Matlab vs. Python) as either argumentative, indicating a preference (better or worse), or non-argumentative. ampere-2019 (Hua et al., 2019) contains 10,386 propositions from 400 peer reviews in machine learning and natural language processing venues, sourced via OpenReview. Following (Park et al., 2015), it encodes the presence or absence of the argumentative discourse units: evaluation, request, fact, reference, or quote. asrd-2020 (Shnarch et al., 2020) contains 700 argumentative sentences from transcribed debate speeches on 20 controversial topics (e.g., sex education, drinking age) as part of IBM Project Debater, where argumentative sentences either support or contest a topic. qmc-2018 (Levy et al., 2018) contains 2,500 argumentative sentences from Wikipedia covering 50 topics (e.g., atheism, early irish astrology), labeled using weak supervision based on signals such as the word 'that' to indicate claims, along with other phrases or concepts denoting argumentative elements. webis-2016 contains 16,402 argumentative segments from 14 domains (e.g., politics, education) sourced from idebate. The data was labeled in accordance with the portal's structure, with users directly responsible for assigning the labels.

Claim-Premise. abstrct-2020, sourced from PubMed, comprises 1,749 sequences extracted from abstracts related to five diseases (e.g., neoplasm, glaucoma), each labeled in the role of either a premise or a claim. amsr-2021 (Fromm et al., 2021), similar to ampere-2019, offers 1,401 labeled sentences from OpenReview peer reviews across six computer science conferences (e.g., human-computer interaction, medical imaging), indicating arguments for or against paper acceptance, or non-argument. finarg-2022 (Alhamzeh et al., 2022) comprises 9,238 sequences labeled as either premise or claim, sourced from transcribed earnings calls of the four major companies Amazon, Apple, Microsoft, and Facebook.

Custom. aec-2015 (Swanson et al., 2015) contains 5,375 sentences on four topics (e.g., gay marriage, gun control) from the now-defunct platforms 4forums and createdebate, highlighting simple argument signals with labels based on the five phrases 'so', 'if', 'but', 'first', and 'I agree that'. taco-2023 (Feger and Dietze, 2024b) consists of 1,734 tweets across six topics (e.g., abortion, Twitter takeover), explicitly designed for cross-topic argument mining on Twitter, focusing on encoding argument components of inference and information while minimizing the topic-related bias, with definitions derived from the Cambridge Dictionary. ukp-2018 (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.

2.4 Textual Features

To numerically represent the datasets, we generated the following basic textual features using spaCy[2].

Syntax is represented by the distribution of POS tags, where tags like SCONJ ('if,' 'because,' 'that') may appear more frequently in arguments.

Length, measured by token and sentence count, as detailed arguments likely need more space.

Readability, measured by the Flesch-Reading-Ease score, indicates text clarity, with lower scores suggesting more complex academic language.

Coherence, by the cosine similarity of consecutive clauses, indicates logical flow at higher values.

Information density is measured by entropy, with higher values indicating greater content variability.

Sentiment, defined by polarity and subjectivity, reveals persuasive strategies through emotions.

Label	Size	Top.	Tok.	Sen.	Rea.	Coh.	Ent.	Pol.	Sub.
ARG	46,526	112	25.70	1.25	62.53	0.08	0.85	0.09	0.39
¬ARG	35,248	114	20.13	1.08	67.96	0.03	0.67	0.08	0.33
Total	81,774	114	23.30	1.17	64.87	0.06	0.77	0.09	0.37

Table 3: Summary of features: topic values are totaled, others averaged, with columns showing the first three letters of the feature names.

We collected 81,774 data points across 114 topics from 11 datasets, with dataset-specific vectors based on mean textual features showing a cosine similarity of 0.96, indicating overall similarity and supporting their integration into a unified ground truth, refer to Table 3. While POS tag distributions varied minimally, nouns (NOUN), proper nouns (PROPN), and punctuation (PUNCT) showed slightly more variation, suggesting early signs of differing contextual word usage between labels. The data shows that 99% of cases have fewer than 128 tokens, are single sentences, moderately readable (avg. ~65), coherent (avg. 0.06), low in entropy (avg. 0.77), neutral in tone (avg. 0.09), and slightly objective (avg. 0.37). Generally speaking, ARG labels show to have more tokens, are easier to read, and provide higher entropy and subjectivity than ¬ARG.

3 Methodology

Random Seeds. We generated 10 random seeds: 5 for sampling and 5 for training. Each experiment uses a unique pair, ensuring random initializations and likely consistent findings (Henderson et al., 2018; Han et al., 2021; D'Amour et al., 2022).

[2] spacy.io/universe/project/textdescriptives

Sampling. To address data imbalances in distributions, we applied stratified sampling, setting the size to 260 instances per label, the minimum in `asrd-2020`, as shown in Table 1, resulting in 520 instances per sample. We applied each sampling seed to the 11 datasets individually, each generating a fixed 70/15/15 split for training, development, and testing per dataset and experiment. This resulted in 28,600 sampled instances, ~67.7% are unique, and almost near-zero sampling error for textual features based on ground truth distributions, with all topics represented. A pilot trial showed that, while using all or only balancing labels per dataset, isolating size and other effects is more time-consuming with no further insights. Thereby, this approach is fair in repetitions, distributions, and representation, given the highly similar datasets.

Train on one, test on another. For each experiment and model, a matrix is created with rows representing source training data and columns representing test data from target datasets. Each entry shows test results after performing hyper-parameter tuning on the target's development set. The final matrix, averaged across all experiments, shows pairwise results of the best performing model.

Leave one out. In this further experiment, all datasets are used for training except one, which is reserved for testing, with results averaged per model after hyper-parameter tuning. The models must generalize from the remaining heterogeneous datasets to the test set, evaluating their ability to transfer learned patterns and adapt to diverse distributions while expanding the training data. We use this as a reference, expecting robust model performance to reveal consistent use of features.

To what transformers pay attention. In this final ablation experiment, we challenge the reported state-of-the-art performance of transformers on individual datasets in the *Train on one, test on another* experiment to explore whether they rely on content-specific cues, given their known focus on basic syntax, nouns, and co-reference. While maintaining the experimental parameters and setup, we filtered the input, as shown in Table 4, by retaining content-dependent parts of speech, specifically nouns (NOUN), proper nouns (PROPN) like 'Gandalf', punctuation (PUNCT), symbols (SYM), and unclassified elements (X), such as 'r2d2bebop', removing all other proper sentence elements. This approach is designed to contradict the notion of generalizable arguments, as it prioritizes content-specific cues over shallow argument signals.

Label	Stage	Example
ARG	Original	A more effective, better tolerated version of this FUP combination is needed.
	Filtered	, version FUP combination .
¬ARG	Original	Ca consumption during the study was equivalent to 1200 mg/d during MD and 525 mg/d during CD.
	Filtered	study mg/d MD mg/d CD .

Table 4: Example from `abstrct-2020` showing the input at different stages (original, filtered) for the *To what transformers pay attention* experiment, keeping only context-specific parts while removing other elements.

Random. We used the dummy classifier of scikit-learn[3], which randomly selects labels, as a baseline to ensure any performance reflects real learning.

Transformers. We selected state-of-the-art transformers BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), DistilBERT (Sanh et al., 2019), and WRAPresentations (Feger and Dietze, 2024a), as other models referenced in the literature were not fully reproducible. WRAPresentations, built on BERTweet (Nguyen et al., 2020) and available on Huggingface[4], is designed for encoding structural argument components while excluding tweet-specific features. Its pre-training makes it particularly interesting for investigating generalization across datasets. Given that parameters are often unspecified in the literature, we use the standard hyper-parameter grid for GLUE text classification, as accepted in the BERT paper and adopted in the RoBERTa paper, balancing performance, time and cost with a batch size of 32, 3 training epochs, and a learning rate between 2e-5 and 5e-5, with each model trained exclusively on a single A-100 GPU.

Decision Tree. We implemented a basic decision tree using scikit-learn[3], leveraging the same pre-computed textual features to represent the ground truth data as a simple opponent to the more complex transformers. Variations included using single features and all k-best features combined. Training on a CPU involved a hyper-parameter grid with tree depths of 1-5, Gini, Entropy, and Log-Loss criteria, and k-best values ranging from 1-7 or all features.

Evaluation. Furthermore, we conducted the hyper-parameter search for the macro F1 score to ensure that the models focus equally on both arguments and non-arguments, emphasizing the importance of treating both classes equally.

[3] scikit-learn.org/stable/
[4] huggingface.co

4 Findings

This section presents the findings, with fine-tuned WRAPresentations referred to as WRAP and the decision tree with all features called the combined decision tree. As outlined earlier, we conducted experiments across various settings, averaging the results over multiple random initializations. In *Train on one, test on another*, transformers perform strongly on individual datasets but mostly struggle to generalize, while simpler decision trees and WRAP generalize better. *Leave one out* results confirm this, emphasizing their robustness against larger and varied training data. The *To what transformers pay attention* experiment reveals that transformers, aside from WRAP, focus primarily on word-label ties specific to the content of individual datasets.

4.1 Train on one, test on another

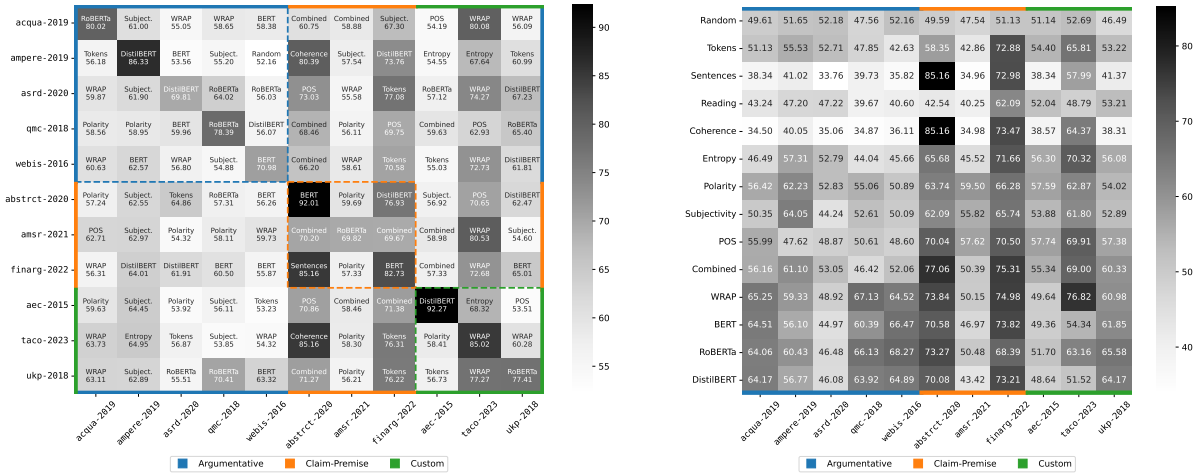
For the 121 pairwise comparisons, WRAP had the best results in 17% of cases, followed by subjectivity (12%) and polarity (11%). BERT and the combined decision tree each led in 10% of cases, with others below this level, refer to Figure 1a.

In the average performance, WRAP achieved the highest macro F1 score of 56.93%, closely followed by the combined decision tree at 56.88%. Similarly, decision trees using POS, subjectivity, and polarity features ranked next with scores between 56.19% and 55.62%, while RoBERTa placed sixth with 55.16%. All others scored lower.

The majority of results, about 71%, fall below a 65% macro F1 threshold in terms of generalization across datasets, largely driven by the decision tree model using various feature sets. This weak generalization can be traced in *Argumentative*, with no scores above 65% macro F1 and the lowest optimum of 52.16%, due to small random deviations.

An analysis of the main diagonal in Figure 1a, which highlights the best performance of models trained on individual datasets, shows that transformer models dominate across all 11 experiments, with macro F1 scores ranging from 69.81% to 92.27%. RoBERTa performed best in four cases, while BERT and DistilBERT led in three each, with WRAP excelling only on the `taco-2023` dataset possibly due to its focus on Twitter data.

In cross-dataset experiments, however, WRAP and the decision tree show their strengths. Despite all classifiers performing below 57% on average, these two models consistently uphold strongest performance and occasionally excel across datasets.



(a) *Train on one, test on another.*

(b) *Leave one out.*

Figure 1: Averaged test results for *(a) Train on one, test on another* and *(b) Leave one out* after hyper-parameter tuning. Independent seeds were used for the 5 stratified samples and classifier training. Both plots show target datasets on the x-axis: in *(a)*, the y-axis represents source datasets in pairwise best-model comparisons, while in *(b)*, classifiers trained on all but the left-out dataset are shown.

4.2 Leave one out

Consistent with prior experiments, the results in Table 1b demonstrate that WRAP achieves the highest performance, with a macro F1 score of 62.87% on average, by being repeatedly trained on all datasets except the one held out for testing, reinforcing its superior generalization across diverse datasets with varying distributions and signals for training.

In contrast, RoBERTa outperformed the combined decision tree in these experiments, with a slightly better macro F1 of 61.63% compared to 59.66%. Moreover, BERT and DistilBERT marginally outperformed classifiers based on features like POS, subjectivity, and polarity, with both models achieving around 59% macro F1. In line with previous experiments, these transformers exceeded the performance of the remaining models.

The comparatively higher macro F1 scores observed for the transformer models, excluding WRAP, may be attributed to their ability to leverage 10 out of 11 datasets for training, thus benefiting from a larger pool of data for training. Despite slight improvements, performance remains low, and transformers do not always perform best.

Across 11 experiments, WRAP delivered the top performance in three instances (27%), while RoBERTa and the combined decision tree each achieved the highest results in two cases (18%). Those decision trees based on length, polarity, subjectivity, and POS each secured the top performance in one experiment (9%) respectively.

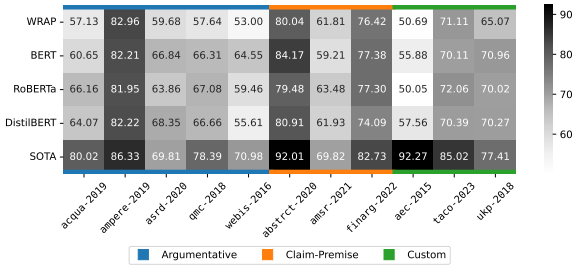


Figure 2: *To what transformers pay attention.* Ablation experiment on *Train on one, test on another* reducing the text to content-specific words, performing close to the state-of-the-art (SOTA). While full data use yields SOTA results, this shows transformers approaching similar performance by relying on contextual words tied to the labels, even without any further sentence elements.

4.3 To what transformers pay attention

The ablation experiment on individual datasets, where transformers were trained solely on words specific to the content of datasets, shows that most models perform close to the referenced state-of-the-art (SOTA) levels achieved in the *Train on one, test on another* experiment, as seen in Tables 1a and 2.

In general, the results approach or exceed 65% macro F1, with the exception of `aec-2015`. For `aec-2015`, the introduction of five argument keywords used for labeling the data resulted in a significant performance drop of 34.71% between the best outcomes with filtered input and the SOTA of the same experiment with proper input, with DistilBERT being the best in both cases.

These findings reveal two trends: First, the datasets acqua-2019, qmc-2018, aec-2015, and taco-2023 show a macro F1 gap of at least 11.31% from SOTA. In contrast, the remaining datasets exhibit smaller differences, ranging from 1.46% for asrd-2020 to 7.84% for abstrct-2020.

Furthermore, all transformers, except WRAP, perform over 68% macro F1 on average. WRAP, with a lower average of 65% macro F1, only excels in the case of ampere-2019, despite being developed for taco-2020.

5 Discussion

One of the most striking findings is the poor generalization, rarely exceeding 65% macro F1, while transformers achieve up to 92.27% macro F1 on individual datasets, as seen in Table 1a. Even across datasets using same definitions, generalization remains inconsistent, as highlighted with 75% of cross-dataset evaluations in *Argumentative* score below 60% macro F1 and none above 65%.

It is important to note that poor generalization is neither due to sub-optimal hyper-parameters, as transformers achieve over 70% macro F1 in some cross-dataset cases, nor to over-fitting, as our pilot study with a larger balanced sample supports these findings, which we elaborate on in the limitations.

Yet, the decision tree with simple features outperforms transformers in 59% of cross-dataset cases, despite lacking advanced representations and fine-tuning. Although, transformers occasionally outperform decision trees, the advantage is minimal and does not significantly enhance generalization, remaining well below isolated benchmark results. Especially in abstrct-2020, BERT achieves a macro F1 of 92.01%, but across all other datasets, generalization is only seen for decision trees, using features like length and coherence, consistently scoring up to 85.16%, as shown in Figures 1a, 1b.

Ablation experiments with minimal input (nouns, proper nouns, punctuation, symbols) show that the strong performance of most transformers on isolated benchmarks, driven by attention to basic syntax, context, and co-reference, mostly results in low difference around 8% from full-sentence benchmarks, with exceptions like asrd-2020 where the gap is as small as 1.46%, as shown in Figure 2.

With an average overlap of 16.31% across all datasets at minimal input, these findings hint that poor transformer generalization reflects content-specific learning of word-label associations.

Nonetheless, the datasets *acqua-2019*, *qmc-2018*, *aec-2015*, and *taco-2023* show differences over 11.31%, refer to Table 2. This is most evident in aec-2015, which is annotated with five argument keywords, where performance falls 34.71% from its 92.27% macro F1 benchmark, showing that transformers can capture simple argument signals if highlighted in the data, but that oversimplified features can also hinder generalizability.

Unlike these datasets, taco-2024 demands experts to focus on labeling shallow structural argument components that generalize across Twitter, while ukp-2018 remains content-specific despite its cross-topic setup for heterogeneous sources.

Results of taco-2024 emphasize the importance of proper methodology in argument identification, as fine-tuning WRAPresentations shows strong generalization among most datasets, achieving up to 80.53% macro F1 on taco-2024 when trained on amsr-2021, compared to the benchmark of 85.02%. WRAPresentations' extended pre-training likely contributes to its success by contrasting argument components of augmented tweets that mimic general text, allowing the model to learn more general representations based on semantics while minimizing susceptibility to content-specific cues. Taken together, this highlights the need for unified, yet not oversimplified, definitions alongside human expertise, proper methods for argument representation and identification while also showing Twitter's role as hub for global debates in argument mining.

6 Conclusion & Ongoing Work

We are the first to question whether current research reflects genuine progress in argument mining, particularly in terms of generalization. Across 11 relevant datasets, we found the transformers BERT, RoBERTa, and DistilBERT, while achieving state-of-the-art performance on isolated datasets, to primarily learn content-specific word-label associations. As a result, these models were outperformed in cross-dataset generalization by a shallow decision tree with simple linguistic features and fine-tuned WRAPresentations, underscoring Twitter's potential as a microcosm of debate, the value of contrastive learning and text augmentation for improving argument generalization. Our findings suggest that simple features and explicit annotations of generalizing signals can guide transformers, highlighting the need to further address differences in definitions and standardize annotation efforts.

Limitations

For our work, we report the following limitations:

While we used publicly available datasets, some (`ukp-2018`, `taco-2024`) require granted access. Additionally, some datasets provided less data than reported, and format inconsistencies (`finarg-2022` in .json, `abstrct-2020` in .ann) complicated integration. For reproducibility, we include a script[1] to retrieve datasets and assemble the ground truth. Some baseline models mentioned in the literature could not be used due to reproducibility or accessibility issues. Nonetheless, we compared our results with the reported benchmarks. We note that this comparison is not entirely proper due to differences in data splits and model selection, but our results were numerically close or better.

Data size may limit these experiments. To balance data size, label distributions, trials, and runtime, samples were made representative and largely independent of random initializations. For comparison, the pilot trails for *Train on one, test on another* and *Leave one out* were run with one seed on a larger sample (68% of the total) by balancing each dataset only by its least frequent label, taking 24 hours for the transformers on an costly A100 GPU, while overall yielding the same trends as reported. On the full dataset, the same experiments took 34 hours for one seed, but imbalances made the results difficult to interpret. The pilot experiments and results are in the repository[1]. In contrast, the three reported experiments ran in 26 hours for all models across five trials, with minimal sampling error and strong data representation.

Although we have primarily focused on addressing content-specific biases within the datasets, we emphasize that differences in definitions and inconsistencies among them are significant issues in their own right. These problems may also hinder the generalization of models trained on these datasets and therefore require further evaluation and research beyond the scope of this work. Our findings should be considered with the understanding that word-label ties within datasets might not be the only reason for poor generalization. Although our results show that we can get very close to state-of-the-art results if using content-specific inputs, there are still differences between these and the results of *To what transformers pay attention*. For datasets that exhibit low influence from content-specific biases, our experiments provide no additional insights into other biases that may limit generalizability.

Acknowledgments

We are deeply grateful for attentive and constructive feedback from our anonymous reviewers, whose expertise and insights are appreciated.

References

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016. 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.

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.

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.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. 2022. Underspecification presents challenges for credibility in modern machine learning. *J. Mach. Learn. Res.*, 23(1).

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.

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2019. Corpus wide argument mining - a working solution. *CoRR*, abs/1911.10763.

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.

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

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2024. Mining legal arguments in court decisions. *Artificial Intelligence and Law*, 32(3):1–38.

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

Wenjuan Han, Bo Pang, and Ying Nian Wu. 2021. Robust transfer learning with pretrained language models through adapters. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 854–861, Online. 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.

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.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. Deep reinforcement learning that matters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

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.

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.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

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

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

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.

Ran Levy, Shai Gretz, Benjamin Sznajder, Shay Hummel, Ranit Aharonov, and Noam Slonim. 2017. Unsupervised corpus–wide claim detection. In *Proceedings of the 4th Workshop on Argument Mining*, pages

79–84, Copenhagen, Denmark. 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.

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

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.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ICAIL '07, page 225–230, New York, NY, USA. Association for Computing Machinery.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

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).

Joonsuk Park, Cheryl Blake, and Claire Cardie. 2015. Toward machine-assisted participation in erulemaking: an argumentation model of evaluability. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, ICAIL '15, page 206–210, New York, NY, USA. Association for Computing Machinery.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1384–1394, San Diego, California. 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.

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.

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2018. Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Melbourne, Australia. Association for Computational Linguistics.

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, 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.

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, 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.

Dusan Varis and Ondřej Bojar. 2021. Sequence length is a domain: Length-based overfitting in transformer models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8246–8257, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691, Osaka, Japan. The COLING 2016 Organizing Committee.

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).

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

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.