A Neural Approach to Fine-Grained Argumentation Strategy Classification with Emotion and Moral Value Lexicons across Multiple Domains

Mohammad Yeghaneh Abkenar[1,3], Weixing Wang[2,3], Manfred Stede[3] and Julia Romberg[4]
[1]Innovation Department, Bundesdruckerei Gruppe GmbH Berlin [2]Hasso Plattner Institute
[3]University of Potsdam [4]GESIS - Leibniz Institute for the Social Sciences
yeghanehabkenar@uni-potsdam.de, weixing.wang@hpi.de,
stede@uni-potsdam.de, julia.romberg@gesis.org

Abstract

Fine-grained argumentation mining goes beyond coarse-grained distinctions such as claim and premise, by delving deeper into the underlying strategies employed (e.g., the use of facts or values to persuade the audience). Despite the advancements brought about by pre-trained language models, the task remains challenging. We investigate whether auxiliary knowledge such as emotion and moral value lexicon features can improve the classification of fine-grained argumentation strategies. Our Neural Flair Transformer Classifier (NFTC), in its base form, fine-tunes a transformer-based document encoder (RoBERTa) for end-to-end argument component classification. Evaluated across four corpora from diverse domains spanning public participation, persuasive forums, product reviews, and student essays, NFTC consistently outperforms majority-voting and Qwen2.5-7B baselines, achieving competitive performance on all datasets. Moreover, gains are observed against a fine-tuned LLaMA-3-8B-Instruct model, regarded in prior work as a leading approach. Injecting additional knowledge into NFTC yields mixed effects: emotion and moral value features provide consistent gains in product reviews and persuasive forums, but not in the other two domains. Our findings suggest that the utility of subjective knowledge is domain and schema dependent, and that knowledge enrichment beyond standard pre-training can meaningfully complement transformer-based models for fine-grained argumentation mining. We provide all resources—including code, the preprocessed corpora, and model architecture—to enable other researchers to build upon our work.[1]

1 Introduction

One of the key tasks in argument mining is the *computational assessment of the function of argumentative units* in natural text. Traditionally, this task has centered on coarse-grained distinctions such as claim (a statement that expresses a specific point or conclusion) and premise (a statement that provides support or justification for a claim) (Palau and Moens, 2009; Liebeck et al., 2016; Stab and Gurevych, 2017; Daxenberger et al., 2017).

While claim/premise classification has been fundamental for argument mining, many applications can benefit from a *more fine-grained analysis that reveals the underlying strategies employed to make claims and premises persuasive for the audience* (Park and Cardie, 2014; Hidey et al., 2017; Dusmanu et al., 2017; Park and Cardie, 2018; Schaefer et al., 2023). Examples include the analysis of persuasion strategies in news editorials (Al-Khatib et al., 2016), or recommending arguments that follow a specific argumentation strategy, which can be used, for example, in debates (Rinott et al., 2015).

The more recent approaches to computationally classifying fine-grained argumentation strategies have put a strong emphasis on the use of pre-trained language models that are subsequently fine-tuned on task-specific data (Schaefer and Stede, 2022; Schaefer et al., 2023; Cabessa et al., 2025). This shift has certainly improved performance, but seems not yet sufficient to accommodate the complex nature of argumentation.

One reason might be the need for additional knowledge beyond the information encoded in the model during the pre-training and fine-tuning phases. Such knowledge can be "any kind of normative information that is considered to be relevant for solving a task at hand and that is not given as task input itself" (Lauscher et al., 2022). In other computational argumentation tasks, we can already observe that such knowledge enrichment can provide further performance improvements, such as in uncovering implicit information (Becker et al., 2020), in audience-specific claim generation (Alshomary et al., 2021), or in stance detection (Abkenar et al., 2026).

[1]The code for the experiments can be found here: **FineAM**.

Proceedings of the 13th Workshop on Argument Mining and Reasoning, pages 94–106
July 3, 2026 ©2026 Association for Computational Linguistics

Our main research question in this paper is therefore **whether, and to what extent, additional knowledge can improve the classification of fine-grained argumentation strategies**. We focus on two sources of information. Our first hypothesis is that *emotionality* can provide useful additional information, as emotional appeal is considered a potential factor affecting the rhetorical effectiveness of arguments (Wachsmuth et al., 2017; Vecchi et al., 2021). Our second hypothesis is that the notion of *moral values* can also provide useful additional information, since references to shared values or moral principles can be effective in persuading audiences in argumentative discourse (Alshomary and Wachsmuth, 2021; Kiesel et al., 2022).

Recent work has shown that argumentation mining models tend to rely on shortcuts and corpora-specific features rather than learning generalizable argumentation properties or subtask, stressing the need to evaluate across different domains (Feger et al., 2025). Motivated by this finding, we test our knowledge-enriched neural models across multiple domains to examine whether lexicon-based features (emotion and moral values) contribute to more robust and generalizable fine-grained argumentation strategy classification.

Our contributions are as follows:

- We evaluate the impact of emotionality signals for fine-grained argumentation strategy classification, using a knowledge-enriched transformer architecture that integrates emotion lexicons (Abkenar et al., 2026).
- We extend the model architecture towards the incorporation of moral value signals.
- We evaluate the approach across four heterogeneous datasets spanning public participation, persuasive forums, product reviews, and essays.
- We find that incorporating emotion and moral value features improves model performance in half of the cases, while leading to a decrease in fine-grained argumentation strategy classification performance for the other half.

2 Related Work

While coarse-grained classification (e.g., into claim and premise) has been a major focus area in argument mining, the more fine-grained look at the argumentation strategies that authors pursue has received comparatively less attention (Schaefer et al., 2023). This is unfortunate, as fine-grained labels often provide richer and more detailed information about argumentative structures.

2.1 Fine-grained Argumentation Schemas

What exactly constitutes a fine-grained argumentation strategy is interpreted diversely and to some extent depends on the goal of a study as well as the domain of the data. Al-Khatib et al. (2016, 2017) target persuasion strategies in news editorials using a schema of *common ground*, *assumption*, *testimony*, *statistics*, *anecdote*, and *other*. In the domain of persuasive essay writing, Carlile et al. (2018) categorize argumentation strategies into the claim types *fact*, *value* and *policy*, which is in line with distinctions made in argumentation theory (cf. (Eggs, 2000)). As premise types, they use *common knowledge*, *real example*, *invented instance*, *warrant*, *statistics*, *testimony*, *definition* and *analogy*. As shown in Figure 1, Schaefer et al. (2023) modify this schema with respect to the premise types, resulting in *testimony*, *statistics*, *hypothetical-instance*, *real-example* and *common-ground*.

To inform a debating agent with Wikipedia knowledge, Rinott et al. (2015) extracts premise types as *study*, *expert*, and *anecdote*. More interactive data sources have also been explored. Dusmanu et al. (2017) and Schaefer and Stede (2022) annotate tweets, either as *factual* or *opinionated*, or by classifying a claim as *unverifiable* or *verifiable* and an evidence as *reason*, *external* or *internal*. Park and Cardie (2018) evaluate fine-grained argumentation strategies (*policy*, *value*, *fact*, *testimony*, *reference*) with the goal of measuring the evaluability of arguments in the context of online public participation platforms. The same schema has also been applied to assess the helpfulness of product reviews (Chen et al., 2022). Park and Cardie (2014) operate on online public participation discussions as well, providing us with a different approach that emphasizes the factuality of propositions. Claims are distinguished as *unverifiable*, *verifiable non-experiential*, and *verifiable experiential*, while premise strategy types are *reason*, *evidence*, and *optional evidence*. Other discussion forum sources include idebate.org, distinguishing supporting arguments as *study*, *factual*, *opinion*, and *reasoning* (Hua and Wang, 2017), and Change My View: (Hidey et al., 2017) use a rather unique

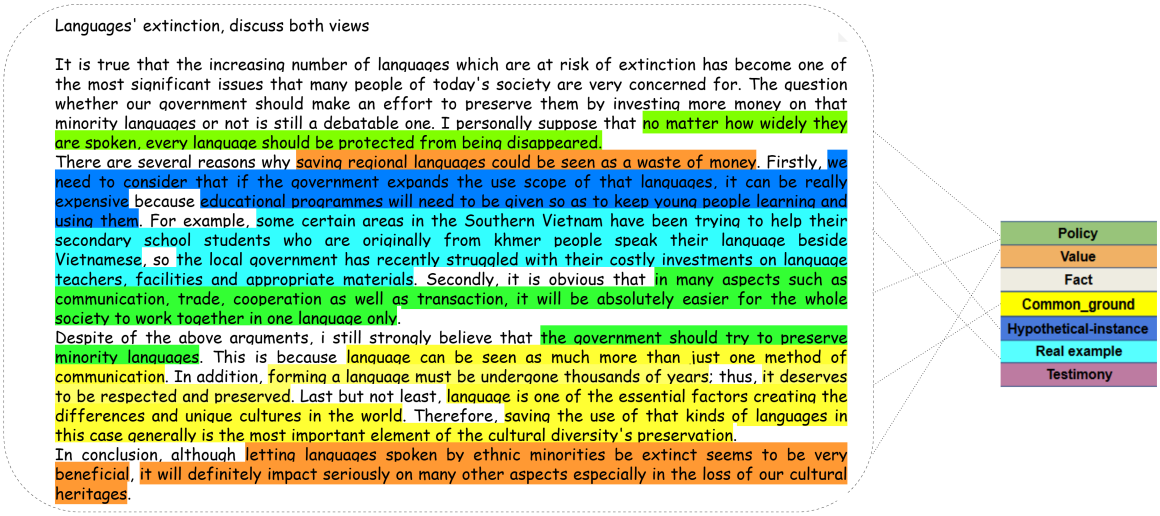


Figure 1: Example of a fine-grained argumentation strategy schema, taken from Schaefer et al. (2023) (AAE-Ext corpus): an argumentative essay annotated with five out of a total of seven fine-grained semantic types (policy, value, fact, common ground, hypothetical instance, real example, testimony). Some labels do not appear in this essay but occur in others.

set of claim types, namely *interpretation*, *evaluation*, *agreement*, and *disagreement*. For premises, Hidey et al. make use of Aristotle's *logos*, *pathos*, and *ethos*, as do a few other works (Habernal and Gurevych, 2017; Carlile et al., 2018).

2.2 Neural Approaches for Fine-grained Argumentation Strategy Classification

When developing approaches to computationally predicting fine-grained argumentation strategies, recent works have demonstrated the dominance of neural methods. Typically, these approaches rely on existing pre-trained language models that are subsequently fine-tuned on task-specific data, such as BERT and RoBERTa (Schaefer and Stede, 2022; Schaefer et al., 2023). In addition, large language models such as LLaMA-3, Gemma-2, Mistral, Phi-3, and Qwen-2 have been utilized for this purpose (Cabessa et al., 2025). A somewhat distinct approach is employed by Chen et al. (2022), who extract argumentative features as input for neural models. Their feature extraction follows Morio et al. (2020)'s method of feeding word features (surface, part-of-speech tags, GloVe and ELMo vectors) into a BiLSTM to predict strategy types.

While these methods achieve promising results, it has been argued that computational models for argumentation require the integration of additional knowledge to adequately capture the complexity and subjective nature of argumentative discourse (Lauscher et al., 2022). We try to fill this gap to a certain extent by revisiting some previously collected datasets, performing preprocessing to prepare corpora suitable for neural methods, and leveraging the injection of **additional knowledge about emotionality and moral values**.

More broadly, fine-grained argumentation mining provides a promising setting for exploring neural models and capturing subtle argumentative phenomena that coarse-grained frameworks may overlook (Ren et al., 2025).

2.3 Emotionality and Moral Values as Signals for Argumentation Strategies

Subjectivity is an integral part of argumentation, with expressions of sentiment, emotion, and affect serving as important signals (Lauscher et al., 2022). Early work provides evidence that such features can be helpful in fine-grained argumentation strategy classification (Rinott et al., 2015; Levy et al., 2014; Dusmanu et al., 2017; Hua and Wang, 2017).

Emotional appeal can affect the rhetorical effectiveness of arguments (Wachsmuth et al., 2017; Vecchi et al., 2021), which is central to the fine-grained study of argumentation properties that make claims and premises persuasive. In particular, research on argument quality often identifies emotion as a key dimension of relevance (Benlamine et al., 2015, 2017; Ziegenbein et al., 2023; Greschner and Klinger, 2025; Quensel et al., 2025; Chen et al., 2026). Pre-neural approaches on fine-grained argumentation strategy classification thus regularly incorporated sentiment and emotion features (Levy et al., 2014; Dusmanu et al., 2017; Hua

and Wang, 2017), in contrast to neural approaches.

References to shared values or moral principles are equally subjective cues (van der Meer et al., 2023; Homayounirad et al., 2025) that can be used as persuasive strategies in arguments (Bench-Capon, 2003). Among the first that connected human values with argument mining was Kiesel et al. (2022), focusing on predicting the values behind arguments. Further studies followed up on this (Jafari et al., 2024; Zhang et al., 2024; Senthilkumar et al., 2025) or explored the role of values in argument generation and rewriting (Alshomary and Wachsmuth, 2021; Shahid et al., 2026). To the best of our knowledge, moral values have not yet been exploited to inform our task at hand.

In sum, these insights motivate our goal to **complement promising neural methods with the subjective signals of emotionality and moral values**.

3 Methodology

We adopt the classification framework recently introduced by Abkenar et al. (2026) for stance classification, combining transformer-based text representations with additional knowledge-driven features. The framework allows external signals to be incorporated alongside contextual embeddings, which makes it suitable for our research question.

3.1 Model Architecture

We first detail the original model architecture, which incorporates emotion features. We then outline how we adapt the architecture in order to include moral value features.

Let $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ denote an argumentative text sequence input consisting of n tokens. The goal of the model is to predict a fine-grained argumentation label $y \in \mathcal{Y}$, where $\mathcal{Y}$ represents the set of dataset-specific fine-grained argumentation categories.

Transformer Encoder. We encode the input text using a pre-trained transformer-based document encoder. The input sequence is formatted as

`[CLS] Argument:` $\mathbf{x}$ `[SEP]`

The transformer produces contextualized token representations

$$\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n\} \in \mathbb{R}^{n \times d},$$

where d is the hidden dimensionality of the encoder. The representation corresponding to the `[CLS]` token, $\mathbf{h}_{\text{CLS}} \in \mathbb{R}^d$, is used as the document-level representation of the argument.

The transformer encoder is fine-tuned end-to-end during training. All models are implemented using the Flair NLP framework[2] with HuggingFace transformer encoders, which we refer to as the Neural Flair Transformer Classifier (NFTC).

Lexicon- and Knowledge-based Features. To enrich the semantic representation with psychologically grounded signals, we incorporate two types of auxiliary features derived from external resources.

- **Emotion features (eNRC).** Emotion information is extracted using the best-performing variant of the extended NRC Emotion Lexicon (eNRC)[3], as introduced by Abkenar et al. (2026). The eNRC expands prior versions of the NRC emotion lexicon (Zad et al., 2021; Mohammad and Turney, 2013) by incorporating additional emotive and affective vocabulary not previously captured. Each lexicon entry is associated with zero or more emotion labels based on Plutchik's Wheel of Emotions (Plutchik, 2001), one of the most influential models in emotion research. This model captures eight primary emotions: *joy*, *trust*, *fear*, *surprise*, *sadness*, *disgust*, *anger*, and *anticipation*.

  For each input text, this produces an emotion feature vector

  $$\mathbf{f}_{\text{emo}} \in \mathbb{R}^8.$$

  Emotion features are constructed by aggregating token-level eNRC scores across the input text, where each emotion category frequency is normalized by the total number of emotionally matched tokens, yielding an 8-dimensional document-level feature vector.

- **Moral foundation features (MoralBERT).** We additionally model moral framing using predictions from MoralBERT[4] classifiers trained on the Moral Foundations Theory (Haidt et al., 2017; Graham et al., 2011). This theoretical framework assumes that human moral reasoning is based on a set of innate, universal moral domains (so-called "foundations"). It organizes morality into ten core

[2] https://github.com/flairNLP/flair
[3] https://github.com/Pioannid/eNRC/tree/main
[4] https://github.com/vjosapreniqi/MoralBERT

Corpora	Domain	Texts	Units	Fine-grained types
CMV	persuasive forum	121 (threads)	4612	policy, value, fact, testimony, rhetorical statement
CDCP	public participation	731 (comments)	4931	policy, value, fact, testimony, reference
AAE-Ext	students' persuasive essays	402 (essays)	6089	policy, value, fact, testimony, statistics, real-example, hypothetical-instance, common-ground, other
AM^2	product reviews	878 (reviews)	6126	policy, value, fact, testimony, reference

Table 1: Corpus statistics of the four preprocessed corpora, annotated with fine-grained argumentation strategies, that are used in our evaluation.

values: *care*, *harm*, *fairness*, *cheating*, *loyalty*, *betrayal*, *authority*, *subversion*, *purity*, and *degradation*.

For each input, MoralBERT outputs a 10-dimensional feature vector,

$$\mathbf{f}_{\text{moral}} \in \mathbb{R}^{10},$$

where each dimension corresponds to one moral foundation category. The value of each entry represents the model-estimated presence or strength of the corresponding moral foundation in the text, normalized to the interval $(0, 1)$. Higher values indicate stronger evidence that the text expresses the corresponding moral foundation.

Feature Integration. The auxiliary features are integrated with the contextual representation via feature-wise concatenation. Given the transformer document embedding $\mathbf{h}_{\text{CLS}}$, the final representation is defined as

$$\mathbf{h}_{\text{final}} = \begin{cases} \mathbf{h}_{\text{CLS}} & \text{(no auxiliary features)} \\ [\mathbf{h}_{\text{CLS}}; \mathbf{f}_{\text{emo}}] & \text{(emotion features)} \\ [\mathbf{h}_{\text{CLS}}; \mathbf{f}_{\text{moral}}] & \text{(moral features)} \end{cases} \quad (1)$$

where $[\cdot;\cdot]$ denotes vector concatenation. Following the approach of Bravo-Marquez et al. (2019), and Abkenar et al. (2026), we have done a simple concatenation of features to be able to compare the effect of both features on the results.

Classification. The resulting document representation $\mathbf{h}_{\text{final}}$ is passed to a linear classification layer that predicts the label distribution over $\mathcal{Y}$. The model is trained using standard cross-entropy loss.

3.2 Corpora and Statistics

In our evaluation, we focus on four English datasets from different domains. These corpora were selected from established argumentation datasets and subjected to extensive pre-processing steps:

- **ChangeMyView (CMV)** (Morio et al., 2019).[5] ChangeMyView is a Reddit subforum dedicated to changing users' views through persuasive argumentation. The corresponding dataset contains 4612 discussion turns, each annotated as *policy*, *value*, *fact*, *testimony*, or *rhetorical statement*. The CMV corpus is provided in Brat[6] annotation format. Each thread contains three posts; an original post, a positive reply, and a negative reply and both inner-post and inter-post relations from which we extract argumentation units and their semantical components types.

- **Cornell eRulemaking Corpus (CDCP)** (Park and Cardie, 2018).[7] CDPC was among the first datasets to adopt a more fine-grained annotation schema, aiming to model argumentative structures as they occur in real-world scenarios. Collected from a US public participation effort, the dataset classifies 4931 argumentative units into *policy*, *value*, *fact*, *testimony*, and *reference*. The CDCP corpus is provided in JSONL format. We extract proposition texts and their semantic components types.

[5] https://katfuji.lab.tuat.ac.jp/nlp_datasets/
[6] https://brat.nlplab.org/
[7] https://huggingface.co/datasets/DFKI-SLT/cdcp

- **Argument Annotated Essays Corpus (AAE-Ext)** (Schaefer et al., 2023).[8] AAE-Ext builds upon the AAE corpus of persuasive essays (Stab and Gurevych, 2017) by introducing an additional annotation layer. It features the largest set of labels in our collection, including *policy*, *value*, *fact*, *testimony*, *statistics*, *real-example*, *hypothetical-instance*, *common-ground*, and *other*. The AAE-Ext corpus is provided in Brat annotation format and consists of 402 essays. We extract the 6089 argument units and their labels *MajorClaim*, *Claim*, and *Premise*. We then integrate the extended version to extract the fine-grained semantic component types.

- **AMazon Argument Mining Corpus (AM^2)** (Chen et al., 2022).[9] 6126 text units from Amazon reviews were categorized according to a fine-grained schema of *policy*, *value*, *fact*, *testimony*, and *reference*. Each entry contains a review with a list of propositions, each annotated with an identifier, argumentation type, text, reasons, and evidence. We extract the proposition texts and their corresponding type labels as input-output pairs for classification.

Table 1 presents the statistics for each corpus, including domain, size, and fine-grained argumentation strategy types.

3.3 Experimental Setup

NFTC. Motivated by the strong results of *RoBERTa* among other encoder models in argument component classification (Schaefer et al., 2023), we fine-tune *RoBERTa-base* for sequence classification as the transformer encoder used in our NFTC approach. Training is conducted for 10 epochs. We use a fixed learning rate of 5×10^{-5} and a mini-batch size of 64.

We evaluate several model configurations of the proposed NFTC classifier. The base model (referred to as NFTC hereinafter) uses only the transformer encoder. To examine the impact of knowledge-enhanced features, we additionally incorporate emotion features derived by means of the eNRC emotion lexicon. We refer to this model variant as NFTC + eNRC. We also evaluate a variant that integrates moral foundation features predicted by MoralBERT (NFTC + MoralBERT). In all cases, the auxiliary feature vectors are concatenated with the transformer [CLS] representation before classification.

Baselines. We compare our three NFTC variants to various baselines. We run a simple majority-class baseline (MajC) and the zero-shot Qwen2.5-7B (Hui et al., 2024) LLM baseline used by Abkenar et al. (2026) on our four evaluation datasets. For Qwen2.5-7B, we use this prompt: *You are an expert text classification model. Task: Read the input text and assign exactly ONE label from the list. Labels: [...] Instructions: - Output ONLY the label text from the list above. - Do NOT output explanations or anything else. Text: [...] Answer (one label only):* Model temperature was set to 0.0.

We further include previously reported results on these datasets, which are taken verbatim from the respective research papers. For the AM^2 dataset, we refer to Chen et al. (2022), who report the performance of a model based on BiLSTM encoders. For CDCP, we resort to the results from Cao (2023), who introduce AutoAM, a multi-task learning model that integrates BERT with a specialized argumentation attention mechanism. For CMV, we reference the results obtained by the structured SVM model in Galassi et al. (2018). For AAE-Ext, however, a direct comparison with prior baseline results is not possible due to differences in the evaluation methodology.

In addition, we compare our approach to LLama-3-8B-Instruct, which was found to be the excelling LLM in argument component classification (Cabessa et al., 2025). Fine-tuning was performed with LLaMA-Factory in SFT mode using QLoRA. We train for 5 epochs on a batch size of 8 and a learning rate of $5e^{-5}$ is used. We fine-tuned in 4-bit quantized form, following prior evidence that this introduces negligible performance differences (Cabessa et al., 2025).

General Settings. We use a 5-fold cross-validation setting in all experiments. The folds remain the same. We have used a seeding method in our publicly accessible code, so the results are fully reproducible. All neural experiments were conducted on a single NVIDIA RTX 3090 GPU.

All models are evaluated on the four corpora (AM^2, CDCP, CMV, and AAE-Ext). Following prior work, model performance is reported as the mean macro F_1 score across the cross-validation folds, due to class imbalance in several datasets. We also report standard deviation across the five folds.

[8] https://github.com/discourse-lab/arg-essays-semantic-types

[9] https://facultystaff.richmond.edu/~jpark/

	AM^2	CDCP	CMV	AAE-Ext
MajC	0.151 (± 0.001)	0.122 (± 0.005)	0.132 (± 0.001)	0.050 (± 0.022)
Qwen2.5-7B	0.629 (± 0.006)	0.724 (± 0.005)	0.685 (± 0.005)	0.389 (± 0.010)
Further Baselines	0.496	0.846	0.735	-
LLaMA-3-8B-Instruct	0.762 (± 0.000)	0.845 (± 0.000)	0.796 (± 0.000)	0.522 (± 0.005)
NFTC (ours)	0.839 (± 0.008)	**0.856 (± 0.010)**	0.834 (± 0.008)	**0.589 (± 0.010)**
NFTC + eNRC	0.858 (± 0.007)	0.821 (± 0.009)	0.840 (± 0.007)	0.567 (± 0.010)
NFTC + MoralBert	**0.861 (± 0.005)**	0.821 (± 0.006)	**0.849 (± 0.005)**	0.463 (± 0.190)

Table 2: Macro F_1 comparison of our Neural Flair Transformer Classifier (NFTC), with the combined eNRC, and moral features variants, the majority-class baseline (MajC), Qwen2.5-7B, LLaMA-3-8B-Instruct, and prior baselines on the corpora. Prior baseline results are taken from Chen et al. (AM^2), Cao (CDCP), and Galassi et al. (CMV).

4 Results and Discussion

Table 2 presents the performance of the different models on the four datasets.

Across datasets, our proposed NFTC classifier demonstrates strong performance and consistent with respect to the standard deviation. On CDCP and AAE-Ext, the base NFTC model achieves the best results, with macro F_1 scores of 0.856 and 0.589, respectively. On AM^2 and CMV, incorporating MoralBERT features yields the highest performance, achieving scores of 0.861 and 0.849. These results substantially outperform the other models.

4.1 Baseline and LLM as a Judge

The majority-class baseline performs poorly across all datasets, achieving macro F_1 scores between 0.050 and 0.151. This confirms the difficulty of fine-grained argumentation strategy classification, particularly for datasets with substantial class imbalance or a larger number of categories such as AAE-Ext (details on the corpora distribution are provided in Appendix A). All neural network-based models substantially outperform this baseline.

Motivated by these findings, we employ Qwen2.5-7B due to its high performance, specialized reasoning capabilities, and efficiency for deployment. The Qwen2.5-7B model provides stronger performance, but remains consistently below the results achieved by our fine-tuned transformer models.

Overall, our NFTC model consistently surpasses both the majority baseline and the LLM baseline across all datasets, indicating that task-specific fine-tuning remains highly effective for fine-grained argumentation strategy classification.

4.2 Prior Models for Fine-grained Argument Classification

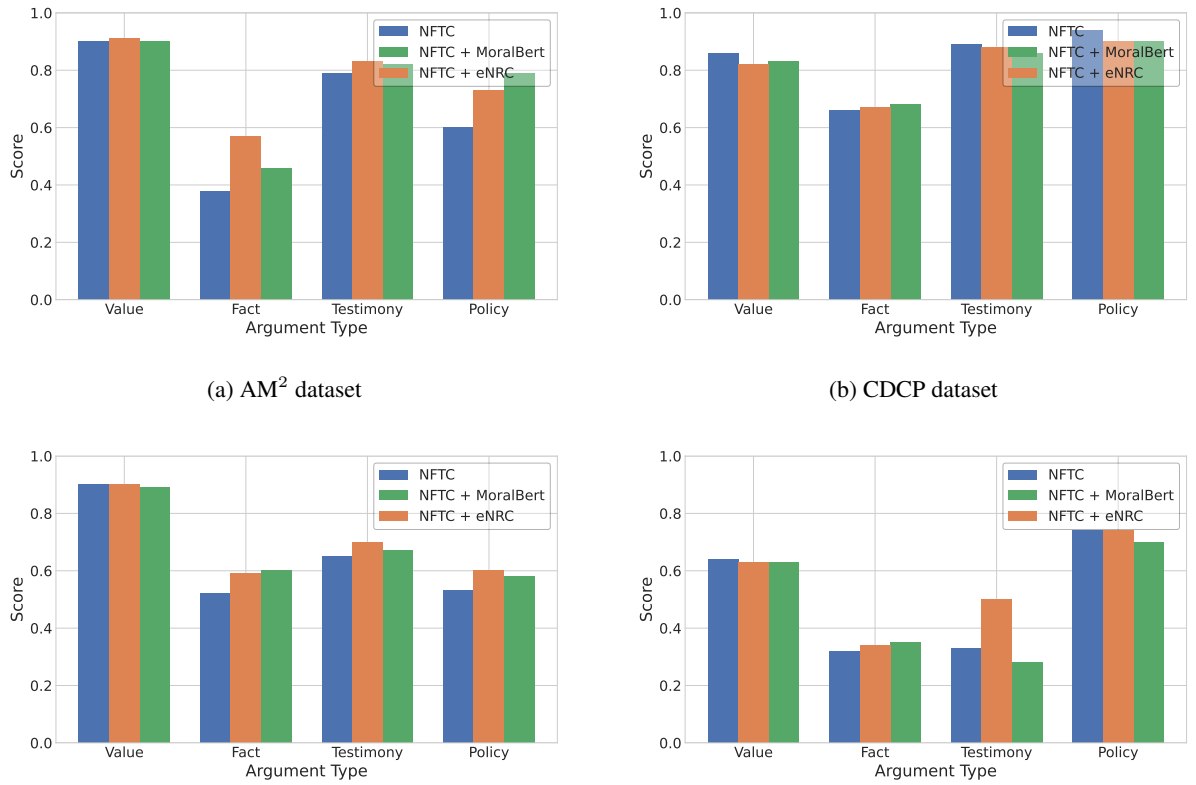
We first compare against results reported on the respective datasets in prior work. As shown in Table 2, for AM^2, NFTC clearly outperforms the previously reported results by Chen et al. (2022). We also obtain better results on CDCP, with 0.856 a slightly higher F_1 score than the previously reported 0.845 of Park and Cardie (2018). For CMV, we achieve improved results with all NFTC variants. Although NFTC yields the best results for AAE-Ext overall, a direct comparison with prior baseline results is not possible due to differences in the evaluation methodology.

Next, we compare our NFTC models to a fine-tuned LLaMA-3-8B-Instruct, which is considered the current state-of-the-art model for argument component classification (Cabessa et al., 2025). While the original authors evaluated their model only on CDCP as a fine-grained argumentation schema, we extend the evaluation to include three additional datasets. Our results demonstrate that even a fine-tuned LLaMA-3-8B-Instruct model does not outperform NFTC variants: for AM^2 and CMV, NFTC consistently achieves better performance, and for CDCP[10] and AAE-Ext, at least one NFTC variant surpasses LLaMA-3-8B-Instruct.

4.3 Effect of Knowledge-Enriched Features

NFTC variants perform best on all datasets. We further evaluate the effect of incorporating external

[10] Cabessa et al. (2025) report a higher mean macro F_1 score of 0.873, compared to our experimental result. However, their evaluation was conducted on a predefined single test split, while we employ a 5-fold cross-validation setup. We argue that our approach provides a more robust estimate of model performance, as it mitigates the risk of results being skewed by the selection of a particular test set.



(a) AM^2 dataset

(b) CDCP dataset

(c) CMV dataset

(d) AAE-Ext dataset

Figure 2: Breakdown of the macro F_1 scores for NFTC in its base form and for the injection of additional knowledge, moral values (NFTC + MoralBERT) and emotion (NFT + eNRC). We report the performance for the overlapping categories, i.e., value, fact, testimony and policy, for all four datasets.

knowledge through emotion features (eNRC) and moral foundation features (MoralBERT). The results indicate that the usefulness of these auxiliary features is dataset-dependent. MoralBERT features improve performance on AM^2 and CMV, suggesting that moral framing contributes to identifying argumentation strategies in product reviews and online discussions. Emotion features also provide improvements, but these are smaller on AM^2 and CMV compared to the NFTC base model.

Moreover, both auxiliary knowledge enrichments lead to a performance drop on the other two datasets, CDCP and AAE-Ext. This finding is particularly interesting when comparing CMV and CDCP, which might be expected to be more similar: one being a general discussion forum and the other focused on political topics. However, a notable difference is likely the target audience: while CMV consists of discussions among users, public participation submissions in CDCP are more directly addressed to officials. For AAE-Ext, the results may also reflect the broader and more heterogeneous annotation schema used in this dataset, where lexical signals alone may provide limited additional information.

Knowledge-based features can complement transformer representations, but their effectiveness depends on the domain, the function of the text, and, presumably, the annotation schema.

4.4 Category-Level Analysis

In order to gain further insights into the impact of emotion and moral value enrichment on the indiviual argumentation strategy categories, Figure 2 presents the category-level performance of the four dataset. The results show that the *value*, *testimony* and *policy* categories are classified with relatively high macro F_1 scores across all model variants, while the *fact* category exhibits lower scores overall.

Notably, the integration of emotion and moral features consistently improves the classification of *facts* across all datasets, and to some extent also enhances the prediction of the *testimony* and *policy* categories. This suggests that these strategies may rely more strongly on affective or normative cues

captured by the auxiliary lexicon-based features. In the case of the objectively normed category of *facts*, a possible explanation is that the absence of emotional or moral value cues provides informative signals to the model. The only category that remains unaffected by model enrichment is *values*, which is particularly interesting given that this category typically reflects highly subjective propositions. Future research is needed to determine whether this is due to a mismatch between the theoretical foundations of our auxiliary features and the *values* category, or if other factors might be contributing to this effect.

5 Conclusion

In this paper, we investigated whether auxiliary knowledge about emotionality and moral values can improve fine-grained argumentation strategy classification. We evaluated our Neural Flair Transformer Classifier (NFTC) across four diverse corpora spanning public participation, persuasive forums, product reviews, and student essays. Our results demonstrate that NFTC consistently outperforms both the majority-voting baseline and Qwen2.5-7B, and achieves competitive performance on all datasets where prior results are available. Furthermore, gains are noted against the fine-tuned LLaMA-3-8B-Instruct model.

The injection of auxiliary knowledge yields mixed effects: MoralBERT features provide consistent gains on AM2 and CMV, while emotion features via eNRC prove more beneficial for AAE-Ext. This suggests that the utility of subjective knowledge is domain and schema dependent, with moral framing being more informative in interactive discourse settings and emotion signals better suited to richer, multi-class annotation schemes. Taken together, our findings support the hypothesis that knowledge enrichment beyond standard pre-training and fine-tuning can meaningfully complement transformer-based models for fine-grained argumentation mining.

Future work should explore more targeted integration strategies — such as attention-based feature fusion (Dai et al., 2021) or gated feature fusion (Li et al., 2020) — and investigate whether other sources of normative knowledge can further close the gap in harder, multi-class settings such as AAE-Ext, as well as extending this approach to relation classification.

Limitations

While our NFTC approach works well from a performance perspective, the improvement brought by the auxiliary features is small. Future work must explore the benefits of auxiliary knowledge in more depth, such as through more complex feature integration methods as well different emotion lexicon features such as SenticNet (Cambria et al., 2016). Additionally, the prompt used for our Qwen2.5-7B baseline was relatively generic and could have been optimized. The results may be further improved by employing prompt engineering approaches.

Ethical Considerations

We are fully aware that systems based on emotion recognition and sentiment analysis can be facilitators of enormous progress, but also enablers of great harm. We therefore strongly advice user of such systems to follow established instructions and ethical guidelines, such as ethics sheets (Mohammad, 2022) and data-sheets for datasets (Gebru et al., 2021). Moreover, emotional expression varies significantly across cultures, ethnic groups, and demographics, as well as moral values. This must be carefully considered, especially when using such systems to support policy-making (i.e., the evalution of public participation).

Acknowledgments

We thank our colleagues in Innovations department of the Bundesdruckerei GmbH, Hasso Plattner Institute and Leibniz Institute for the Social Sciences (GESIS) for providing us with the opportunity to freely work on our research topics. Thank you for fostering an environment that encourages innovation and academic growth. The authors also sincerely thank the anonymous reviewers for their thoughtful recommendations that significantly improved this paper.

References

Mohammad Yeghaneh Abkenar, Weixing Wang, Manfred Stede, Mark A. Finlayson, Davide Picca, and Panagiotis Ioannidis. 2026. Improving neural argumentative stance classification in controversial topics with emotion-lexicon features. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 9678–9691, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. 2017. Patterns of argumentation strategies across topics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1351–1357, Copenhagen, Denmark. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. 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.

Milad Alshomary, Wei-Fan Chen, Timon Gurcke, and Henning Wachsmuth. 2021. Belief-based generation of argumentative claims. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 224–233, Online. Association for Computational Linguistics.

Milad Alshomary and Henning Wachsmuth. 2021. Toward audience-aware argument generation. *Patterns*, 2(6):100253.

Maria Becker, Ioana Hulpuş, Juri Opitz, Debjit Paul, Jonathan Kobbe, Heiner Stuckenschmidt, and Anette Frank. 2020. Explaining arguments with background knowledge. *Datenbank Spektrum*, 20:131–141.

Trevor J. M. Bench-Capon. 2003. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448.

Mohamed S. Benlamine, Serena Villata, Ramla Ghali, Claude Frasson, Fabien Gandon, and Elena Cabrio. 2017. Persuasive argumentation and emotions: An empirical evaluation with users. In *Human-Computer Interaction. User Interface Design, Development and Multimodality*, pages 659–671, Cham. Springer International Publishing.

Sahbi Benlamine, Maher Chaouachi, Serena Villata, Elena Cabrio, Claude Frasson, and Fabien Gandon. 2015. Emotions in Argumentation: an Empirical Evaluation. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015*, Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, pages 156–163, Buenos Aires, Argentina.

Felipe Bravo-Marquez, Eibe Frank, Bernhard Pfahringer, and Saif M. Mohammad. 2019. Affectivetweets: a weka package for analyzing affect in tweets. *Journal of Machine Learning Research*, 20(92):1–6.

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.

Erik Cambria, Soujanya Poria, Rajiv Bajpai, and Björn Schuller. 2016. Senticnet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, pages 2666–2677.

Lang Cao. 2023. Autoam: An end-to-end neural model for automatic and universal argument mining. In *International Conference on Advanced Data Mining and Applications*, pages 517–531. Springer.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631, Melbourne, Australia. Association for Computational Linguistics.

Yanran Chen, Lynn Greschner, Roman Klinger, Michael Klenk, and Steffen Eger. 2026. Emotionally charged, logically blurred: AI-driven emotional framing impairs human fallacy detection. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6709–6732, Rabat, Morocco. Association for Computational Linguistics.

Zaiqian Chen, Daniel Verdi do Amarante, Jenna Donaldson, Yohan Jo, and Joonsuk Park. 2022. Argument mining for review helpfulness prediction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8914–8922, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. 2021. Attentional feature fusion. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3560–3569.

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.

Mihai Dusmanu, Elena Cabrio, and Serena Villata. 2017. Argument mining on Twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2317–2322, Copenhagen, Denmark. Association for Computational Linguistics.

Ekkehard Eggs. 2000. Vertextungsmuster Argumentation: Logische Grundlagen. In Klaus Brinker, editor, *Text- und Gesprächslinguistik*, volume 16 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, pages 397–414. Walter de Gruyter, Berlin.

Marc Feger, Katarina Boland, and Stefan Dietze. 2025. Limited generalizability in argument mining: State-of-the-art models learn datasets, not arguments. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23900–23915, Vienna, Austria. Association for Computational Linguistics.

Andrea Galassi, Marco Lippi, and Paolo Torroni. 2018. Argumentative link prediction using residual networks and multi-objective learning. In *Proceedings of the 5th Workshop on Argument Mining*, pages 1–10.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. 2011. Mapping the moral domain. *Journal of personality and social psychology*, 101(2):366.

Lynn Greschner and Roman Klinger. 2025. Fearful falcons and angry llamas: Emotion category annotations of arguments by humans and LLMs. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 628–646, Albuquerque, USA. Association for Computational Linguistics.

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

Jonathan Haidt, P Ditto, J Graham, R Iyer, S Koleva, M Motyl, G Sherman, and S Wojcik. 2017. Moral foundations theory. *Social Theorists of Morality*, page 261.

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.

Amir Homayounirad, Enrico Liscio, Tong Wang, Catholijn M Jonker, and Luciano Cavalcante Siebert. 2025. Will annotators disagree? identifying subjectivity in value-laden arguments. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15237–15252, Suzhou, China. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. 2017. Understanding and detecting supporting arguments of diverse types. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 203–208, Vancouver, Canada. Association for Computational Linguistics.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Amir Reza Jafari, Praboda Rajapaksha, Reza Farahbakhsh, Guanlin Li, and Noel Crespi. 2024. Unveiling human values: Analyzing emotions behind arguments. *Entropy*, 26(4).

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.

Anne Lauscher, Henning Wachsmuth, Iryna Gurevych, and Goran Glavaš. 2022. Scientia potentia est—on the role of knowledge in computational argumentation. *Transactions of the Association for Computational Linguistics*, 10:1392–1422.

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.

Xiangtai Li, Houlong Zhao, Lei Han, Yunhai Tong, Shaohua Tan, and Kuiyuan Yang. 2020. Gated fully fusion for semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11418–11425.

Matthias Liebeck, Katharina Esau, and Stefan Conrad. 2016. What to do with an airport? mining arguments in the German online participation project tempelhofer feld. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 144–153, Berlin, Germany. Association for Computational Linguistics.

Saif Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Gaku Morio, Ryo Egawa, and Katsuhide Fujita. 2019. Revealing and predicting online persuasion strategy with elementary units. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6274–6279, Hong Kong, China. Association for Computational Linguistics.

Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, Yuta Koreeda, and Kohsuke Yanai. 2020. Towards better non-tree argument mining: Proposition-level biaffine parsing with task-specific parameterization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3259–3266, Online. Association for Computational Linguistics.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, page 98–107, 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.

Joonsuk Park and Claire Cardie. 2018. A corpus of eRulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Carlotta Quensel, Neele Falk, and Gabriella Lapesa. 2025. Investigating subjective factors of argument strength: Storytelling, emotions, and hedging. In *Proceedings of the 12th Argument mining Workshop*, pages 126–139, Vienna, Austria. Association for Computational Linguistics.

Yupei Ren, Xinyi Zhou, Ning Zhang, Shangqing Zhao, Man Lan, and Xiaopeng Bai. 2025. Towards comprehensive argument analysis in education: Dataset, tasks, and method. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14215–14231.

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.

Robin Schaefer, René Knaebel, and Manfred Stede. 2023. Towards fine-grained argumentation strategy analysis in persuasive essays. In *Proceedings of the 10th Workshop on Argument Mining*, pages 76–88, Singapore. Association for Computational Linguistics.

Robin Schaefer and Manfred Stede. 2022. GerCCT: An annotated corpus for mining arguments in German tweets on climate change. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6121–6130, Marseille, France. European Language Resources Association.

Rithik Appachi Senthilkumar, Amir Homayounirad, and Luciano Cavalcante Siebert. 2025. Leveraging large language models to identify the values behind arguments. In *Value Engineering in Artificial Intelligence*, pages 87–103, Cham. Springer Nature Switzerland.

Farhana Shahid, Stella Zhang, and Aditya Vashistha. 2026. Llms homogenize values in constructive arguments on value-laden topics. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA. Association for Computing Machinery.

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

Michiel van der Meer, Piek Vossen, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2023. Do differences in values influence disagreements in online discussions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15986–16008, Singapore. Association for Computational Linguistics.

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.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Samira Zad, Joshuan Jimenez, and Mark Finlayson. 2021. Hell hath no fury? correcting bias in the NRC emotion lexicon. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 102–113, Online. Association for Computational Linguistics.

He Zhang, Alina Landowska, and Katarzyna Budzynska. 2024. Detection and analysis of moral values in argumentation. In *Value Engineering in Artificial Intelligence*, pages 114–141, Cham. Springer Nature Switzerland.

Timon Ziegenbein, Shahbaz Syed, Felix Lange, Martin Potthast, and Henning Wachsmuth. 2023. Modeling

appropriate language in argumentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4344–4363, Toronto, Canada. Association for Computational Linguistics.

A Appendix

Dataset	Label	Count	(%)
AM^2 (6126)	value	3752	61.25
	testimony	1965	32.08
	fact	281	4.59
	policy	124	2.02
	reference	4	0.07
CDCP (4931)	value	2177	44.15
	testimony	1118	22.67
	policy	815	16.53
	fact	789	16.00
	reference	32	0.65
CMV (4727)	value	3134	66.30
	rhetorical_statement	727	15.38
	testimony	354	7.49
	fact	257	5.44
	policy	140	2.96
	major_claim	115	2.43
AAE-Ext (6089)	common_ground	1774	29.13
	value	1502	24.67
	hypothetical_instance	917	15.06
	real_example	717	11.78
	fact	411	6.75
	statistics	400	6.57
	policy	344	5.65
	testimony	22	0.36
	other	2	0.03

Table 3: Label distribution across datasets. Counts and percentages are reported for each class.