Towards Fine-Grained Argumentation Strategy Analysis in Persuasive Essays

Robin Schaefer and **René Knaebel** and **Manfred Stede**
Applied Computational Linguistics
University of Potsdam
14476 Potsdam, Germany
{robin.schaefer|rene.knaebel|stede}@uni-potsdam.de

Abstract

We define an *argumentation strategy* as the set of rhetorical and stylistic means that authors employ to produce an effective, and often persuasive, text. First computational accounts of such strategies have been relatively coarse-grained, while in our work we aim to move to a more detailed analysis. We extend the annotations of the Argument Annotated Essays corpus (Stab and Gurevych, 2017) with specific types of claims and premises, propose a model for their automatic identification and show first results, and then we discuss usage patterns that emerge with respect to the essay structure, the "flows" of argument component types, the claim-premise constellations, the role of the essay prompt type, and that of the individual author.

1 Introduction

The field of Argument Mining (AM), which has grown into a productive area of research during the last decade (Stede and Schneider, 2018; Lawrence and Reed, 2020), focuses on the tasks of automatic identification and extraction of argumentation in natural language. This includes the detection of argument components, like claims (Daxenberger et al., 2017; Schaefer et al., 2022) and premises (Rinott et al., 2015), and the relations between them (Carstens and Toni, 2015). Research has been conducted on different text domains ranging from more edited texts, e.g. editorials (Al-Khatib et al., 2016) or Wikipedia texts (Rinott et al., 2015), to social media, e.g. Change My View (Hidey et al., 2017) or Twitter (Schaefer and Stede, 2022).

A so far relatively understudied area of interest is the identification of argumentation strategies, i.e., the decisions that authors make on linearizing their argumentation and on marking it with linguistic expressions for persuasive effect (Al-Khatib et al., 2017; El Baff et al., 2019). Effectiveness, which can be described as one dimension of argumentation quality (Wachsmuth et al., 2017), depends (inter alia) on using the "right" arguments for the audience, their arrangement, and their linguistic formulation. This is also the case for persuasive essays, which already have been extensively used in AM research (Stab and Gurevych, 2014b; Wachsmuth et al., 2016), but to the best of our knowledge not much for modeling underlying strategies. To contribute to filling this gap we utilize our own claim and premise type annotations to extract semantic "flow patterns" from the Argument Annotated Essays (AAE) corpus (Stab and Gurevych, 2017). We argue that these types and flow patterns are fine-grained and informative to shed more light on the strategies authors of persuasive essays apply to structure their texts.

The contributions of this paper are: 1) We provide a dataset with claim and premise *type* annotations for the full AAE corpus (Sct. 3) by revising and extending the prior work of Carlile et al. (2018); 2) we trained classification models on our annotations and present first promising results (Sct. 4); 3) we contribute to argumentation strategy modeling by (i) extracting flow patterns of the argument component types, also in relation to the essay structure (roles of different paragraphs), (ii) examining the patterns of claim and supporting premise w.r.t. their types, and (iii) looking into the influences of essay prompt as well as the individual author of the text (Sct. 5).

2 Related Work

Argument Mining in Essays. Stab and Gurevych (2014a) presented the first edition of the AAE corpus, which consisted of 90 persuasive essays annotated for argument components and relations. Later, it was extended to 402 essays (Stab and Gurevych, 2017). This corpus has been repeatedly used for component detection (Stab and Gurevych, 2014b; Schaefer et al., 2022) and as a starting point for component type

Proceedings of the 10th Workshop on Argument Mining, pages 76–88

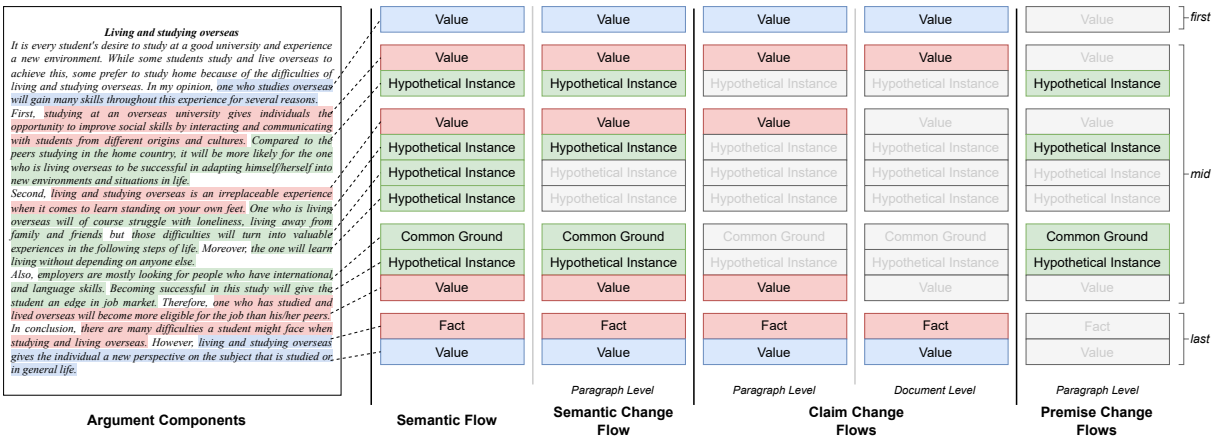

Figure 1: Overview: Essay #5 (Stab and Gurevych, 2017) with argument component types: major claim (blue), claim (red), and premise (green). Based on our semantic types, different variants of semantic flows are demonstrated.

annotations (Carlile et al., 2018).

Considerable work has been dedicated to automated essay quality scoring (Ke and Ng, 2019). While essays often were assigned only holistic scores, more recently research has shifted towards the investigation of individual dimensions of essay quality, e.g., coherence or persuasiveness (Nguyen and Litman, 2018). While argument patterns and strategies are related to essay quality, in this paper we do not specifically investigate the implications for quality but leave that to future work.

Argument Component Types. Type tagsets have been proposed for different argument components and data domains. In more formal texts like Wikipedia articles (Rinott et al., 2015), news editorials (Al-Khatib et al., 2016) or argumentative essays (Carlile et al., 2018) premises are categorized as *study/statistics*, *expert/testimony*, *anecdote* and/or *common knowledge/common ground*, among others. Hua and Wang (2017) annotated the types *study*, *factual*, *opinion*, and *reasoning* in idebate.org texts. With respect to claims, Carlile et al. (2018) assigned the types *fact*, *value*, and *policy*, as well as Aristotle's modes of persuasion *logos*, *pathos*, and *ethos* (Higgins and Walker, 2012). Recently, Chen et al. (2022) annotated argumentative units in Amazon reviews with the types *fact*, *testimony*, *policy*, and *value* in order to enable review helpfulness prediction.[1]

For Twitter, Addawood and Bashir (2016) applied a set of premise types to *news media accounts*, *blog posts*, or *pictures*. Dusmanu et al. (2017) annotated argumentative tweets according to them being *factual* or *opinionated*. More recently, Schaefer and Stede (2022) used the premise types *reason* and *external/internal evidence* and annotated claims for their *un/verifiability* (Park and Cardie, 2014). Other relevant social media platforms include the subreddit Change My View. Hidey et al. (2017) assign a rather unique set of types to claims, consisting of *interpretation*, *evaluation-rational*, *evaluation-emotional*, and *agreement/disagreement*, while annotating premises with *logos*, *pathos*, and *ethos*.

In our work, we use a modified set of claim and premise types for annotation, which has been derived from the annotation guidelines applied by Al-Khatib et al. (2016) and Carlile et al. (2018).

Argument Patterns. Wachsmuth et al. (2016) experiment with argumentative discourse unit (ADU) flows. They train models on argumentative essays in AAE (Stab and Gurevych, 2014a) to automatically identify argument components in the larger ICLE corpus (Granger et al., 2020). In contrast to their work, we use more fine-grained semantic classes instead of the argument component types themselves. We expect more informative patterns for describing the writing strategies in student essays. Al-Khatib et al. (2017) adapt previous work, extract evidence types (*statistics*, *testimony*, *anecdote*) in argumentative newspaper editorials, and show differences across automatically classified topics.

3 Corpus & Annotation

In this section, we briefly describe the corpus we use, i.e. the AAE corpus (Stab and Gurevych, 2017). In addition, we present our annotation scheme, the procedure, and results.

[1]While their vocabulary overlaps with Carlile et al. (2018), their definitions (except for *policy*) are notably different.

	Examples	Type
1)	[...] we should attach more importance to cooperation during primary education.	P
2)	[...] keeping the cultural traditions in the destination countries is tremendous important.	V
3)	[...] teachers teach us knowledge but also the skills to tell right from wrong.	F
4)	Frank Zappa once said, "Mind is like a parachute, it doesn't work if its not open"	T
5)	The waste products and harmful gases produced by these factories cause a significant amount of air pollution.	S
6)	[...] if there are no animals in the world, the balance of nature will broke down, and we, human, will die out as well.	HI
7)	[...] tourism makes up one-third of the Czech Republic's economy.	RE
8)	Nowadays, time is the most valuable thing in life with increased pace.	CG

Table 1: Examples of semantic type annotations. Abbreviations: P (policy), V (value), F (fact), T (testimony), S (statistics), HI (hypothetical-instance), RE (real-example), CG (common-ground). Linguistic errors in the original text have not been corrected.

3.1 The Argument Annotated Essays Corpus

The AAE corpus (Stab and Gurevych, 2017) consists of 402 persuasive student essays, which were written in response to a prompt (e.g. *International tourism is now more common than ever before. Some feel that this is a positive trend, [...]. What are your opinions on this?*). The essays have been annotated for the core components of argumentation, i.e., (major) claim and premise. Persuasive essays tend to exhibit a rather rigid structure, which is reflected in the actual usage of the components.

An essay starts with an introduction, which usually contains the *major claim*. The major claim is the author's main stance regarding the essay's topic, i.e., the prompt. The introduction is followed by several paragraphs in which the actual argumentation unfolds. Each paragraph contains one or more arguments, consisting of a *claim* and at least one *premise*, the latter of which supports or attacks the former. The claim bears a stance toward the major claim. Thus, a unit's argument role depends on its position in the argumentative tree; e.g., a unit directly relating to a major claim is a claim.

In this work, we add another annotation layer to the corpus, claim types and premise types. While Carlile et al. (2018) annotated semantic types for only 102 essays, we applied our modified annotation scheme to the full corpus of 402 essays.

3.2 Annotation Scheme

We derived and modified our annotation scheme from the schemes created by Carlile et al. (2018) and Al-Khatib et al. (2016). We annotate three claim types (*policy*, *value* and *fact*) and five premise types (*testimony*, *statistics*, *hypothetical-instance*, *real-example* and *common-ground*).[2] We motivate our decision to apply a new annotation scheme as follows: 1) In our initial experiments, annotating the dataset using the guidelines by Carlile et al. (2018) was challenging and repeatedly led to low IAA. 2) Some types were rarely annotated (analogy, definition) or difficult to define (warrant). These were removed from our set. 3) Some types were also used in other studies (e.g., testimony and statistics; Al-Khatib et al. (2016)) and allow for a potential comparison across corpora. See Table 1 for annotation examples.

Claim Types. We annotated the same claim types as Carlile et al. (2018) but modified their definitions in order to facilitate the annotation process. All types are defined with a focus on the author's argumentative intention, i.e., what they argue for. As this is usually left somewhat implicit, the annotator needs to take into account the context of the essay to understand the author's reasoning.

A *policy* claim is used to argue towards some action to be taken or not to be taken, while a *value* claim attaches a certain value to a target, e.g., good/bad or important/unimportant. Importantly, this often is achieved using implicit means, which complicates the annotation procedure. Finally, a *fact* claim is used to argue in favor or against some target statement being true or false, i.e., it asserts

[2]Our data and annotation guidelines can be found here: https://github.com/discourse-lab/arg-essays-semantic-types.

something to (not) hold in the world. Crucially, a *fact* claim does not need to state an actual truth in the world (fact checking is a separate issue) but is used to convince the audience of the target's assumed truthfulness or falseness. As these classes semantically overlap to a certain degree, we apply a claim annotation hierarchy: policy > value > fact.

Premise Types. The premise types were initially derived from those of Carlile et al. (2018). However, as testing the guidelines in early annotation sessions did not yield promising results, we refined our guidelines using the evidence type definitions of Al-Khatib et al. (2016).

A *testimony* unit gives evidence by stating or quoting that a proposition was made by an expert, authority, group, or similar. The expert can be explicitly named, but a more general usage is also accepted, such as "Scientists suggest that...". *Statistics* states the results of quantitative research or studies, and also includes more general phrasings that refer to quantitative analyses and dependencies. The latter focuses on proportions, aggregations like the mean, correlations, or similar dimensions.

We apply two *example* categories, viz. *real-example* and *hypothetical-instance*. A *real-example* describes either a real (historical) event, that can be located in space and/or time, or a specific statement about the world. The event or statement can be "proven" by an external source, which does not need to be given. While the author's personal experiences are treated as *real-example*, usually described using 1st person pronouns, statements adopting any 3rd person perspective are treated as *testimony*. A *hypothetical-instance* is similar to a *real-example*, but as it is hypothetical it was conceived merely by the author and thus cannot be verified by an external source.

A *common-ground* unit includes statements being depicted as common knowledge or self-evident fact. In other words, the author presents them as being accepted without evidence by most readers. In contrast with the example categories, *common-ground* refers to general issues, not to specific events or statements. Finally, we use an *other* class to allow for the annotation of units that the annotator is uncertain about. We apply the following premise annotation hierarchy: testimony > statistics > hypothetical-instance > real-example > common-ground > other.

Annotation Class	Krippendorff's α
Policy	0.78
Value	0.52
Fact	0.34
Claim Type	*0.52*
Testimony	-
Statistics	0.16
Hypothetical-Instance	0.70
Real-Example	0.58
Common-Ground	0.42
Other	-
Premise Type	*0.53*

Table 2: Inter-annotator agreement.

3.3 Annotation Procedure

Three annotators, one of whom is a co-author of this paper, were trained to annotate the data. On a paragraph-by-paragraph basis the annotation task consists of 1) annotating the types of all claims and 2) annotating the types of all premises.

Annotators were trained in an iterative manner. A first draft of the guidelines was tested by two annotators in an initial round of 20 essays. Afterward, IAA was calculated, and feedback was given by the annotators leading to revised guidelines. These steps were repeated until acceptable IAA scores were obtained. Then, the third annotator was trained using the final annotation guidelines and another set of 20 essays. Once all annotators were able to complete the task, they labeled the same set of 40 essays, i.e., 10% of the corpus, in order to calculate the final IAA scores. Finally, two annotators continued labeling until the whole corpus was annotated (with one single judgement).

3.4 Annotation Results

We evaluate our annotation guidelines in terms of Krippendorff's α (Artstein and Poesio, 2008). In addition to the IAA by component (claim and premise) we calculate alpha for individual semantic types by using a binary "class vs. not class" distinction. See Table 2 for the IAA.

With respect to claim types, annotators obtained the best results for the policy class (0.78). *Value* yielded a score of 0.52, while the fact class obtained a score of 0.34. Calculating IAA on the set of all claim type annotations received a score of 0.52.

Considering premise type annotation, the best results were obtained for hypothetical-instance (0.70) and real-example (0.58), which are both example classes. Common-Ground achieved a score of 0.42. The statistics class posed a challenge for annota-

Annotation Class	Counts	Proportion
Policy	344	0.15
Value	1,502	0.67
Fact	411	0.18
Testimony	22	0.01
Statistics	400	0.10
Hypothetical-Instance	917	0.24
Real-Example	717	0.19
Common-Ground	1774	0.46
Other	2	0.00

Table 3: Annotation statistics: counts and proportions.

tors (0.16). As our set of 40 essays did not provide enough testimony to calculate IAA, we cannot present results for this class. Altogether, annotators achieved a score of 0.53 for the set of all premise types.

3.5 Corpus Statistics

In this work, we provide another annotation layer for the AAE corpus. Hence, all basic corpus statistics were obtained from the originally published dataset.[3] The corpus consists of 402 essays with a mean token count of 357 (min: 207; max: 550) and a mean sentence count of 17 (min: 8; max: 33). On average the essays consist of five paragraphs (min: 3; max: 7), including the introduction and the final paragraph. The paragraphs have a mean ADU count of 3 (min: 1; max: 12).

Our annotations show a notable class imbalance (see Table 3). *Value* is the dominant claim type with a proportion of 0.67, followed by *fact* (0.18) and *policy* (0.15). With respect to premise types, *common-ground* was annotated most frequently (0.46). The *example* categories *hypothetical-instance* and *real-example* show a comparable proportion (0.24 vs 0.19), while *statistics* has been identified more rarely (0.10). *Testimony* shows a small proportion of 0.01. *Other* only has been annotated twice and will be ignored in the following sections.

4 Classification of Semantic ADU Types

We fine-tune a pre-trained language model, the *roberta-base* architecture (Liu et al., 2019), to predict semantic types. As input we use solely the argument component span, without further context. See Appendix A for details on hyper-parameters. Our complete classification results are also provided there; in the following, we mention the main points.

We train the semantic type classifiers separately for the different ADU types (major claim, claim, premise), and in addition with the variant of combining the two claim types (major claim and claim). Per run, the data is randomly divided into train/dev/test with proportions 80/10/10. The results that we report are averaged over 10 runs.

Previous State of the Art. To allow for comparison with previous research by Ke et al. (2018), we first train our neural model on their originally annotated 102 essays (henceforth referred to as PREVIOUS). While they provide only accuracy (micro-average) results, we will below, in contrast, present a more detailed report with a per-class evaluation. Our accuracy for claim type predictions is better, with 76.9% compared to 69.5% reported by Ke et al. (2018). For premise types, we achieved 70.1% accuracy, compared to 31.2%.

The main contribution to our increase in performance is probably due to the pre-trained language model. A closer look at the premises' macro F1 scores reveals that the only class that is well-predicted is *common-ground* (81.5 F1), followed by *real-example* (65.6 F1) and *statistics* (30.3 F1). Three out of eight classes (*analogy*, *testimony*, and *definition*) have no predictions at all, due to the imbalanced class distribution.

Baseline. As a baseline for the experiments with our own annotations on the full corpus (402 texts), we take the simple prediction of the most frequent (majority) semantic type observed in the training data per ADU type. This yields macro scores for major claims and claims of 26.2 F1 and 26.9 F1, respectively, while for premises it amounts to only 13.5 F1, in part due to the higher number of classes.

Results. Trained on our annotation, the neural model clearly outperforms the baselines. For both major claim types (75.9 F1) and claim types (77.2 F1), we achieved very good results. In comparison to PREVIOUS, our claim predictions increased by 12.4 F1. While we perform better on *value* and *policy* classification, PREVIOUS has higher scores on *fact*, which is probably due to different label distributions: Two-thirds of the claim labels in the data of Carlile et al. (2018) are facts. Unexpectedly, training with a fused class of the two claim types has not led to an improved performance. While the F1 score for *fact* is marginally better, the

[3] We use the Trankit Toolkit (Nguyen et al., 2021) for data preprocessing.

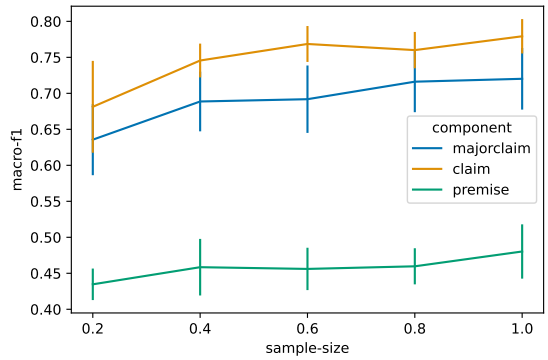

Figure 2: Learning curve with respect to sample size.

performance of the other two classes (*value* and *policy*) does not improve. Here, the results are just in between the separately trained models.

For the premise predictions, we achieve similar performance (70.2%) as PREVIOUS in terms of accuracy. However, the higher macro-average of our model (56.6 F1) compared to PREVIOUS (25.0 F1) indicates a better balance of our per-class predictions. The complete results are shown in Appendix A.

Data Size Learning Curve. We study how our additionally annotated data affects the performance of our neural model. We run the same experiments, but after splitting the test set we use only a subsample for training and development.[4]

Figure 2 shows the varying performance of our models with different portions (20% to 100%) of training data. While the increase for semantic types of premises is not particularly high, a larger increase in performance is evident for the two other ADU classes, claim and major claim. This shows that the effort of additional annotation is justified.

5 Pattern Extraction and Analysis

Argumentative essays have a very specific structure, as described in Section 3. Following previous works on argument component types (Wachsmuth et al., 2016) and argumentation strategies (Al-Khatib et al., 2017), we hypothesize finding similar patterns of semantic argument types in the essays.

First, we linearize the semantic types and order them by their textual positions. In Figure 1, for example, the essay starts with a value-based major claim in the first paragraph, followed by a value claim and a premise with semantic type *hypothetical-instance*. The full sequence of semantic types is referred to as the *semantic flow* of the essay. We further abstract over semantic repetitions, thus resulting in so-called *semantic change flows*. For example, in the previous flow, multiple consecutive *hypothetical-instances* are reduced to a single occurrence. This abstraction leads to more reliable/general patterns (Al-Khatib et al., 2017). Similar to Wachsmuth et al. (2016), we use the natural structure of argumentative essays and split them into paragraphs, as individual arguments are fully contained in single paragraphs. This reduces the length of extracted patterns and their variance.

Additionally, we also study differences in the semantic change flows of component types. For claim change flows, besides paragraphs we study their change flow on full documents, too. As claims should only relate to the major claim, we assume document-level change flows should summarize the global structure of an essay quite well. For premises, we restrict our study to the paragraph level, as premises should not be connected to the premises of other paragraphs.

Argument components show different distributions across paragraphs, with major claims only appearing in the first and last, and premises predominantly being used in the middle paragraphs. This has an effect on the semantic flows. See Table 4 for our semantic change flow results.

Regarding the change flows of claim types (see Table 4 (a)), the first paragraph often only contains flows consisting of a single unit, usually a major claim (value: 0.35; policy: 0.18; fact: 0.07). If a flow of two units can be found, a major claim usually precedes a claim. This pattern deviates from the last paragraph, where the major claim is reformulated. It is common for a change flow to start or end with a major claim. The middle paragraphs are dominated by individual claim types (value: 0.65; fact: 0.23; policy: 0.09), while changes from one type to another occur more rarely. With respect to claim change flows over full essays, changes between types most prominently occur 2-4 times. Usually two major claim types are combined with 1-3 claim types. The value type is most commonly applied, which is reflected by the distribution of type annotations. Individual combinations of value and fact types are a more common pattern than other claim type combinations.

Considering the change flows of premise types (see Table 4 (b)) *common-ground* is the most common type, It is used either as an individual flow or

[4]We make sure that the different component models use the same test split across varying sample ratios.

Level	#	Change Flow	Freq
par-first	1	(M:Value)	0.35
	2	(M:Policy)	0.18
	3	(C:Value)	0.11
	4	(M:Fact)	0.07
	5	(M:Value, C:Value)	0.06
	6	(C:Value, M:Value)	0.04
	7	(C:Fact)	0.03
	8	(M:Policy, C:Value)	0.03
	9	(C:Policy)	0.02
	10	(M:Value, C:Fact)	0.02
par-mid	1	(C:Value)	0.65
	2	(C:Fact)	0.23
	3	(C:Policy)	0.09
	4	(C:Fact, C:Value)	0.01
	5	(C:Value, C:Fact)	0.01
par-last	1	(M:Value)	0.23
	2	(M:Value, C:Value)	0.14
	3	(C:Value, M:Value)	0.14
	4	(M:Policy)	0.08
	5	(M:Policy, C:Value)	0.08
	6	(C:Value, M:Policy)	0.04
	7	(C:Fact, M:Value)	0.03
	8	(M:Fact)	0.03
	9	(M:Value, C:Fact)	0.03
	10	(M:Value, C:Policy)	0.02
full	1	(M:Value, C:Value, M:Value)	0.09
	2	(M:Value, C:Value, M:Value, C:Value)	0.05
	3	(M:Value, C:Value, C:Fact, M:Value)	0.04
	4	(C:Value, M:Value)	0.03
	5	(M:Value, C:Fact, C:Value, M:Value)	0.03
	6	(M:Value, C:Value, C:Fact, C:Value, M:Value)	0.02
	7	(M:Value, C:Fact, C:Value, M:Value, C:Value)	0.01
	8	(M:Policy, C:Value, M:Policy, C:Value)	0.01
	9	(C:Value, C:Fact, M:Value)	0.01
	10	(C:Value, M:Value, C:Value)	0.01

(a) Claim change flows.

Level	#	Change Flow	Freq
par-mid	1	(CG)	0.20
	2	(CG, HI)	0.11
	3	(HI)	0.07
	4	(CG, RE)	0.06
	5	(CG, HI, CG)	0.04
	6	(S, CG)	0.04
	7	(RE)	0.03
	8	(HI, CG)	0.03
	9	(S)	0.03
	10	(CG, S)	0.02

(b) Premise change flows.

Level	#	Change Flow	Freq
par-mid	1	(C:Value, CG)	0.08
	2	(C:Value, HI)	0.04
	3	(C:Value, CG, HI)	0.03
	4	(CG, C:Value)	0.03
	5	(C:Fact, CG)	0.03
	6	(C:Value, RE)	0.02
	7	(C:Value, S, CG)	0.02
	8	(CG, HI, C:Value)	0.02
	9	(C:Value, CG, RE)	0.02
	10	(C:Value, CG, HI, CG)	0.02
	11	(C:Value, HI, CG)	0.01
	12	(C:Policy, CG)	0.01
	13	(CG, C:Value, CG)	0.01
	14	(C:Fact, CG, HI)	0.01
	15	(C:Value, S)	0.01
	16	(CG, C:Fact)	0.01
	17	(CG, RE, C:Value)	0.01
	18	(C:Value, HI, RE)	0.01
	19	(C:Value, CG, HI, CG, HI)	0.01
	20	(C:Value, RE, CG)	0.01

(c) Claim and premise change flows.

Table 4: Most common change flows of semantic types for different argument components. The letters M and C followed by a colon refer to major claim and claim, respectively. For premise types, we use the abbreviations: CG (common-ground), HI (hypothetical-instance), RE (real-example), and S (statistics).

in combination with other types, the latter of which most often starts with *common-ground*. The most prominent change flow consisting of three types is *(CG, HI, CG)*. A combination of the two example types *hypothetical-instance* and *real-example* is not among the most frequent change flows. *Statistics* most often co-occurs with *common-ground*.

Finally, the claim and premise change flows by paragraph (see Table 4 (c)) reveal that a middle paragraph most often begins with a value claim followed by at least one premise of a certain type. More complex change flows contain *common-ground* and *hypothetical-instance* (e.g. (C:Value, CG, HI); (C:Value, CG, HI, CG)). Flows including fact claims are slightly more frequent than flows including policy claims.

Patterns of Claim-Premise Pairs. In addition to the extraction of semantic type flows we are interested in analyzing the patterns of claims with their direct premise dependents (see Table 5). While the former is focusing on linear order, the latter is hierarchical in nature.

All claim types exhibit the same order of types among their direct premise dependents, i.e., *common-ground* is the most dominant type, followed by *hypothetical-instance*, *real-example*, *statistics*, and *testimony*. This order is reflected by annotation proportions. However, differences between claim types can be observed with respect to the distribution of premise types. Policy claims are associated with a notably larger proportion of *common-ground* (0.59 vs. 0.47/0.43) and a smaller proportion of *real-example* (0.11 vs. 0.19/0.17), while also showing the largest difference between *common-ground* and the following premise type *hypothetical-instance*. Fact claims are supported by the largest proportion of *statistics* (0.15 vs. 0.09/0.10). *Hypothetical-instance* is rather equally distributed with a small drop for policy claims.

Effects of Prompt Type and Author. As the argumentative essays were written in response to prompts, we are interested in identifying their po-

Claim Type	Premise Type	Proportion
Policy	Common-Ground	0.59
	Hypothetical-Instance	0.20
	Real-Example	0.11
	Statistics	0.09
	Testimony	0.01
Value	Common-Ground	0.47
	Hypothetical-Instance	0.24
	Real-Example	0.19
	Statistics	0.10
	Testimony	<0.01
Fact	Common-Ground	0.43
	Hypothetical-Instance	0.25
	Real-Example	0.17
	Statistics	0.15
	Testimony	<0.01
	other	<0.01

Table 5: Claim heads and their direct premise dependents. Only support relations are considered.

tential effect on the claim type distribution. To achieve this we annotated each prompt with a type from our set of claim types. As the whole *prompt* can consist of multiple propositions, we only consider its central message in combination with the actual prompting sentence, which is often phrased as a question. Consider the prompt example shown in Section 3.1: *International tourism is now more common than ever before. Some feel that this is a positive trend, [...]. What are your opinions on this?*. While this *prompt* bears some complexity, it primarily asks the author to present their opinion on whether the growth of international tourism represents a positive or negative trend. Thus, this prompt is labeled with type *value*.

After the prompt annotation, we calculated the claim type class distribution by prompt type.[5] Due to duplicates among the prompts we only consider 370 individual prompts in our analysis (see Figure 3). While value claims are dominant across all prompts, it is notable that the prompt type has an effect. Policy prompts elicit essays with a rather high policy claim proportion (0.33) while essays in response to value and fact prompts rarely show *policy*. Furthermore, essays written in response to fact prompts show the highest proportion of fact claims (0.28 vs. 0.15/0.16) while value prompts elicit essays with the highest proportion of value claims (0.77 vs. 0.52/0.68).

Another potential factor of interest is the author, i.e., the usage of argument types may depend on the

[5]Prompt types are distributed as follows: policy: 0.37; value: 0.48; fact: 0.15.

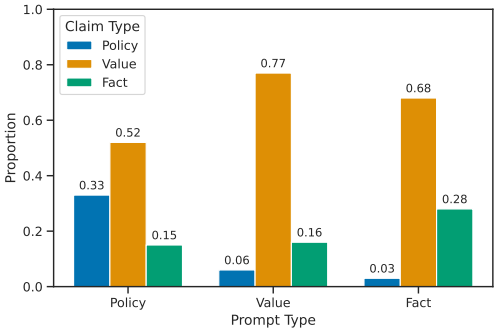

Figure 3: Claim type proportions by prompt type.

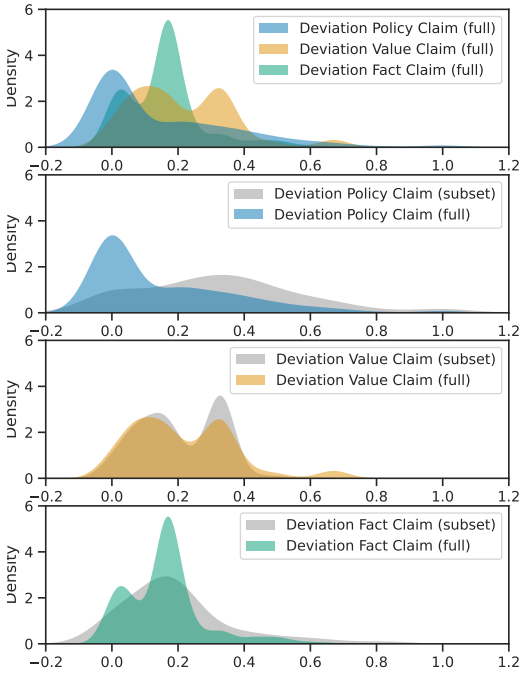

Figure 4: Density plots of absolute deviation from proportion median by claim type for full dataset (plot 1) and by prompt type subset (*policy*: plot 2; *value*: plot 3; *fact*: plot 4).

person writing the essay. In order to investigate this question we calculated each essay's absolute deviation from the median proportion by claim type. The median was calculated using the 370 essays elicited by individual prompts. We use density plots to show the distribution of absolute deviation (see Figure 4, plot 1). The analysis reveals a substantial difference in distribution by claim type. While the deviation from the median of the policy proportion is positively skewed, the deviation of the value annotations shows a bimodal distribution. The fact annotations' deviation also show a bimodal distribution, albeit with a stark difference in density between major and minor modes. While being differently distributed, all claim types show

a notable deviation from the respective proportion median.

While both prompt and author may have an independent effect, they may interact with each other (see Figure 4, plots 2-4; see Appendix B for a full analysis). We show the effect of prompt type by splitting the dataset accordingly and individually comparing the distribution of deviation per claim type with the respective distribution of the full dataset. Plot 2 reveals that the deviation of policy claim annotations in the policy prompt subset is more broadly spread than in the full dataset. In the fact prompt subset, the deviation distribution of fact claim annotations resembles a normal distribution, while it is bimodal in the full dataset (plot 4). However, the distributions of deviation of value claim annotations appear to be similar in both the value prompt subset and the full dataset (plot 3).

6 Discussion

Our change flow analysis reveals several frequently occurring patterns. To begin with, an essay usually starts with a major claim (most frequently of type *value* or *policy*) which is sometimes followed by a claim. The final paragraph, however, shows more flexibility regarding the ordering of both claim variants, which shows that some authors choose to end with a major claim, i.e., the essay's central claim. Moreover, middle paragraphs either contain a single claim (a single argument) or several claims of the same type, which may show an author's tendency to not switch between claim types within a paragraph. Then, while both major claims and claims are most frequently of type *value*, we found a notable difference in the usage of policy and fact types. While *policy* more often occurs in major claim flows, i.e., in the first and last paragraphs, *fact* is more prominently applied as a claim type in the middle paragraphs. Thus, an essay's central claim is more often arguing towards some action being taken, while the argumentation unfolding in the essay's body more often focuses on the question if a target is true or not.

Regarding the usage of premise types we observe the frequent pattern of flows starting with *common-ground* and ending with a different type, or, alternatively, of *common-ground* framing another type. Hence, authors tend to begin their flow of premises with a general statement, followed, e.g., by an example. Less often, *common-ground* is applied to end a flow, while being rarely used in-between types. This may be indicative of a strategy to begin (and end) with a general observation while more concrete statements are placed in-between.

In this work, we explore the effect of two potential factors on the constellation of claim types, prompt type and author, and their potential interaction. Our prompt type analysis provides evidence that the prompt's phrasing has indeed an effect on the usage of claim types, as all prompt types elicit essays with a higher proportion of the respective claim type. Thus, authors adapt their argumentation strategy to the task at hand. We also show that authors exhibit a substantial variance in their usage of claim types, which is further dependent on the essay's prompt type. We argue that this is indicative of the task's role in choosing the most efficient argumentation strategy.

7 Conclusion

In this work, we analyzed patterns of claim and premise types in persuasive essays to shed light on underlying argumentation strategies. We extended the annotations of the AAE corpus with a layer of semantic types, which we used for automatic type classification, pattern extraction both on the level of change flows and argument relations, and the analysis of prompt and author effects on argumentation strategies.

We show that semantic types of argument components are an appropriately fine-grained level of analysis to investigate argumentation strategies. Several common patterns of semantic type flows could be identified. Furthermore, we provide evidence for the effect of author and, especially, prompt type on the adoption of argumentation strategies.

In the future, we would like to extend our scope of analysis. Further research can include the relation between prompt type and semantic flows and the effect of prompt type on the usage of premise types. We are also interested in investigating the effect of semantic flows on essay quality. Finally, we want to apply our analysis to other corpora, both in-domain (the ICLE dataset (Granger et al., 2020)) and out-of-domain (e.g., the subreddit Change My View).

Limitations

In this work, we use a corpus that consists of learner essays that exhibit a rather wide range of language levels. This may influence the distribution of patterns, as presumably the argumentation will be of

different complexity.

Furthermore, while being a standard corpus in AM research, the AAE corpus offers only a limited amount of data. This is reflected in some classes being rarely represented (e.g., testimony) and affects the success of the semantic type classification. Thus, applying the framework to different data such as the ICLE dataset becomes important for getting a better impression of used patterns in persuasive essays.

Further limitations concern our analyses. So far we have not investigated the relation between prompts and semantic flows, which could yield important insights on differences in argument patterns with respect to the task. We also concentrated on the effect of prompt type and author on the usage of claim types, while ignoring their effect on the premise type distribution.

Ethics Statement

Our annotations are based on the publicly available AAE corpus. We point out that information about the essays' authors is not known for this corpus. Thus it is not possible to assess whether these essays are well distributed and representative of a broader audience.

Acknowledgements

This research has been supported by the German Research Foundation (DFG) with grant number 455911521, project "LARGA" in SPP "RATIO". We would like to thank Sophia Rauh and Hugo Meinhof for their annotation efforts, and the anonymous reviewers for their valuable feedback.

References

Aseel Addawood and Masooda Bashir. 2016. "what is your evidence?" a study of controversial topics on social media. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 1–11, Berlin, Germany. Association for Computational Linguistics.

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.

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Comput. Linguist.*, 34(4):555–596.

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.

Lucas Carstens and Francesca Toni. 2015. Towards relation based argumentation mining. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 29–34, Denver, CO. 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.

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.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, Manfred Stede, and Benno Stein. 2019. Computational argumentation synthesis as a language modeling task. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 54–64, Tokyo, Japan. Association for Computational Linguistics.

Sylviane Granger, Maïté Dupont, Fanny Meunier, Hubert Naets, and Magali Paquot. 2020. *International Corpus of Learner English. Version 3*. Presses universitaires de Louvain.

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.

Colin Higgins and Robyn Walker. 2012. Ethos, logos, pathos: Strategies of persuasion in social/environmental reports. *Accounting Forum*, 36(3):194–208. Analyzing the Quality, Meaning and Accountability of Organizational Communication.

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.

Zixuan Ke, Winston Carlile, Nishant Gurrapadi, and Vincent Ng. 2018. Learning to give feedback: Modeling attributes affecting argument persuasiveness in student essays. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4130–4136. International Joint Conferences on Artificial Intelligence Organization.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6300–6308. International Joint Conferences on Artificial Intelligence Organization.

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

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Huy Nguyen and Diane Litman. 2018. Argument mining for improving the automated scoring of persuasive essays. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A lightweight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.

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.

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. 2022. On selecting training corpora for cross-domain claim detection. In *Proceedings of the 9th Workshop on Argument Mining*, pages 181–186, Online and in Gyeongju, Republic of Korea. International Conference on 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.

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

Christian Stab and Iryna Gurevych. 2014b. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

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

Manfred Stede and Jodi Schneider. 2018. *Argumentation Mining*, volume 40 of *Synthesis Lectures in Human Language Technology*. Morgan & Claypool.

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.

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.

A Hyper-Parameters & Experimental Results

For argument component classification, we use RoBERTa (Liu et al., 2019) for sequence classification. In particular, we choose the *roberta-base* architecture implemented by HuggingFace.[6] We freeze the first half of the model and only fine-tune the second half in order to reduce the computational effort.

For optimization, we use AdamW (Loshchilov and Hutter, 2019) with $5e^{-5}$ learning rate. The batch size is set to 16 throughout our experiments. We train for 10 epochs and linearly reduce the learning rate over the number of training steps.

The best model is chosen based on the best average loss during validation. Each component type's samples are randomly divided into three parts, train/dev/test with portions 80/10/10, respectively. All experiments are repeated 10 times and the reported results cover mean and standard deviation.

ADU	Semantic-Type	Precision	Recall	F1
MAJORCLAIM	fact	$0.457_{0.218}$	$0.475_{0.149}$	$0.448_{0.159}$
	value	$0.910_{0.034}$	$0.918_{0.040}$	$0.914_{0.031}$
	policy	$0.937_{0.056}$	$0.898_{0.074}$	$0.915_{0.049}$
	micro avg			$\mathbf{0.881}_{0.043}$
	macro avg	$0.768_{0.079}$	$0.763_{0.065}$	$0.759_{0.049}$
CLAIM	fact	$0.592_{0.069}$	$0.581_{0.089}$	$0.583_{0.067}$
	value	$0.857_{0.037}$	$0.862_{0.032}$	$0.859_{0.023}$
	policy	$0.871_{0.071}$	$0.882_{0.059}$	$0.874_{0.047}$
	micro avg			$\mathbf{0.800}_{0.030}$
	macro avg	$0.773_{0.032}$	$0.775_{0.031}$	$0.772_{0.020}$
(MAJOR-)CLAIM	fact	$0.626_{0.069}$	$0.556_{0.095}$	$0.587_{0.081}$
	value	$0.871_{0.021}$	$0.901_{0.021}$	$0.885_{0.015}$
	policy	$0.893_{0.030}$	$0.874_{0.052}$	$0.882_{0.030}$
	micro avg			$\mathbf{0.836}_{0.016}$
	macro avg	$0.796_{0.025}$	$0.777_{0.026}$	$0.786_{0.020}$
PREMISE	hypothetical-instance	$0.694_{0.065}$	$0.699_{0.052}$	$0.695_{0.049}$
	common-ground	$0.739_{0.025}$	$0.759_{0.034}$	$0.749_{0.024}$
	real-example	$0.785_{0.066}$	$0.731_{0.051}$	$0.756_{0.048}$
	statistics	$0.435_{0.050}$	$0.433_{0.063}$	$0.431_{0.042}$
	testimony	$0.208_{0.315}$	$0.233_{0.335}$	$0.193_{0.264}$
	micro avg			$\mathbf{0.702}_{0.030}$
	macro avg	$0.572_{0.054}$	$0.571_{0.060}$	$0.566_{0.048}$

Table 6: Class specific results (Ours) across argument components and the combination of claims and major claims.

ADU	Semantic-Type	Precision	Recall	F1
CLAIM	fact	$0.848_{0.069}$	$0.872_{0.053}$	$0.857_{0.042}$
	value	$0.584_{0.189}$	$0.556_{0.158}$	$0.538_{0.115}$
	policy	$0.579_{0.293}$	$0.645_{0.380}$	$0.549_{0.288}$
	micro avg			$\mathbf{0.769}_{0.061}$
	macro avg	$0.670_{0.104}$	$0.691_{0.106}$	$0.648_{0.098}$
PREMISE	common-knowledge	$0.744_{0.070}$	$0.911_{0.063}$	$0.815_{0.044}$
	warrant	$0.058_{0.118}$	$0.058_{0.118}$	$0.058_{0.118}$
	invented-instance	$0.250_{0.344}$	$0.154_{0.238}$	$0.164_{0.221}$
	real-example	$0.771_{0.089}$	$0.596_{0.177}$	$0.656_{0.120}$
	analogy	$0.000_{0.000}$	$0.000_{0.000}$	$0.000_{0.000}$
	testimony	$0.000_{0.000}$	$0.000_{0.000}$	$0.000_{0.000}$
	statistics	$0.467_{0.476}$	$0.242_{0.270}$	$0.303_{0.319}$
	definition	$0.000_{0.000}$	$0.000_{0.000}$	$0.000_{0.000}$
	micro avg			$\mathbf{0.701}_{0.055}$
	macro avg	$0.286_{0.087}$	$0.245_{0.066}$	$0.250_{0.065}$

Table 7: Class specific results (PREVIOUS) of our model on the 102 essays annotated by Carlile et al. (2018).

[6] www.huggingface.com

B Density Plots: Effects of Prompt Type and Author

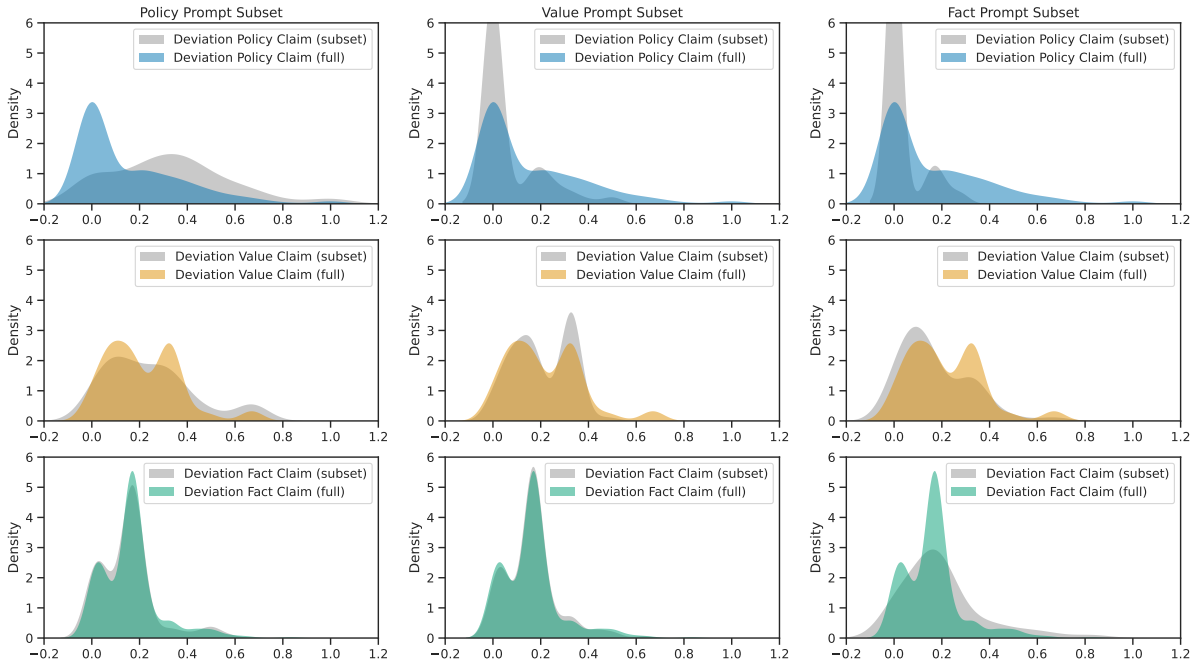

Figure 5: Comparison of density plots of absolute deviation from proportion median by claim type between full dataset and prompt type subsets. The rows are split by claim type. The columns are split by prompt type.