Using Argument Mining to Assess the Argumentation Quality of Essays

Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein
Faculty of Media, Bauhaus-Universität Weimar, Germany
{henning.wachsmuth,khalid.alkhatib,benno.stein}@uni-weimar.de

Abstract

Argument mining aims to determine the argumentative structure of texts. Although it is said to be crucial for future applications such as writing support systems, the benefit of its output has rarely been evaluated. This paper puts the analysis of the output into the focus. In particular, we investigate to what extent the mined structure can be leveraged to assess the argumentation quality of persuasive essays. We find insightful statistical patterns in the structure of essays. From these, we derive novel features that we evaluate in four argumentation-related essay scoring tasks. Our results reveal the benefit of argument mining for assessing argumentation quality. Among others, we improve the state of the art in scoring an essay's organization and its argument strength.

1 Introduction

Argument mining aims to determine the argumentative structure of natural language texts. Usually, this structure is composed of different types of argumentative discourse units, such as premises and conclusions, that together form one or more arguments in favor of or against some thesis.

One of the main proposed downstream applications of argument mining is writing support including automated grading, which will extend the capabilities of massive open online courses (MOOCs), thereby contributing to unlimited access and participation in education. To aid argumentative writing, we envision a writing support system to proceed in three major steps: (1) The *mining* of argumentative structure, (2) the *assessment* of specific quality dimensions based on the mined structure, and (3) the *synthesis* of suggestions for quality improvements. Figure 1 visualizes the resulting process. Several approaches to the mining step have been developed and evaluated in terms of the effectiveness of the mined structure. So far, however, the benefit of this structure remains largely unexplored (see Section 2 for details).

This paper puts the assessment step into the focus. We ask if, to what extent, and how the output of argument mining can be leveraged to assess the argumentation quality of a text. In particular, we consider these questions for persuasive student essays. Such an essay seeks to justify a thesis on a given topic via a composition of arguments. Different quality dimensions related to argumentation have been studied for persuasive essays, such as the clarity of the justified thesis (Persing and Ng, 2013). Also, argument mining has already been performed effectively on persuasive essays (Stab and Gurevych, 2014b).

We build on the outlined research in that we use argument mining to assess an essay's argumentation quality. First, we adapt a state-of-the-art approach for mining argumentative discourse units (Section 3). Then, we apply the approach to all essays from the International Corpus of Learner English (Granger et al., 2009) in order to analyze their argumentative structure. We find statistically reliable patterns that yield insights into how students argue in essays. From these, we derive novel solely structure-oriented features for machine learning (Section 4). Finally, we tackle essay scoring for four argumentation-related quality dimensions: organization, thesis clarity, prompt adherence, and argument strength. In systematic experiments, we compare our features to strong baselines and to the state of the art (Section 5). The observed results provide clear evidence for the impact of argumentative structure on argumentation quality: Our features consistently do best among all structure-oriented approaches. Moreover, we outperform the state of the art of scoring the organization and the argument strength of persuasive essays.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1680–1691, Osaka, Japan, December 11-17 2016.

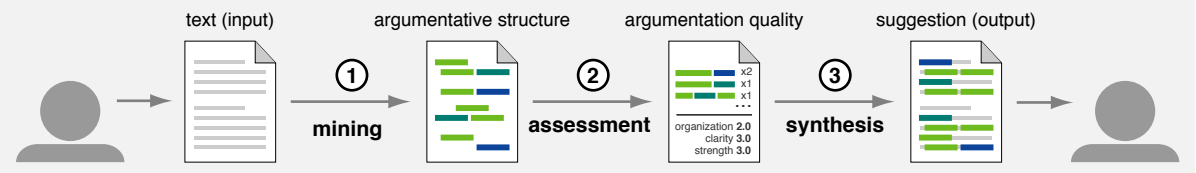

Figure 1: The three major steps of the envisioned process of writing support systems.

Contributions Altogether, with this paper we provide the following contributions to research:

1. We examine the use of argument mining for assessing argumentation quality for the first time.
2. We reveal common patterns in the argumentative structure of persuasive essays statistically.
3. We provide the new state of the art approach to two argumentation-related essay scoring tasks.

2 Related Work

Several approaches to argument mining have been introduced, often grounded in argumentation theory: Matching the argumentation schemes of Walton et al. (2008), Mochales and Moens (2011) model each argument in legal cases as a conclusion with a set of premises. Based on (Freeman, 2011), Peldszus and Stede (2015) capture support and attack relations between argumentative discourse units of microtexts. Habernal and Gurevych (2015) adapt the fine-grained argument model of Toulmin (1958) for web texts. As detailed in Section 3, we rely on the essay-oriented model of Stab and Gurevych (2014a). For us, mining is a preprocessing step only, though. For statistical reliability, we restrict our view to the units of arguments. Like Moens et al. (2007), we classify units on the sentence level, but we consider four different unit types. This results in a sequential structure comparable to argumentative zones (Teufel et al., 2009). The latter have also been exploited for downstream applications (Contractor et al., 2012).

Our focus is the *analysis* of argumentative structure. Related structures have been analyzed before: To measure text coherence, Feng et al. (2014) build on discourse structure (Mann and Thompson, 1988), which is connected but not equivalent to argumentative structure (Peldszus and Stede, 2013). Faulkner (2014) classifies the stance of essays using argument representations derived from dependency parse trees. For essay scoring, Persing et al. (2010) detect the discourse function of each paragraph in an essay in order to align the resulting function sequence with known function sequences. Similarly, we capture a review's overall structure in (Wachsmuth et al., 2014a) by comparing the local sentiment flow in the review to a set of common flow patterns that are learned through clustering. In (Wachsmuth et al., 2015), we further abstract the flows to optimize their domain generality in global sentiment analysis. Discourse structure, discourse functions, and sentiment flows serve as baselines in our experiments in Section 5. Unlike all mentioned approaches, however, we analyze the output of argument mining.

In particular, we use the mined structure to assess argumentation quality. While there is no common definition of such quality, Blair (2012) specifies the goals of relevance, acceptability, and sufficiency for arguments. To find accepted arguments in debate portals, Cabrio and Villata (2012) analyze attack relations between arguments based on the framework of Dung (1995). Rinott et al. (2015) detect three types of evidence in Wikipedia articles, and Boltužić and Šnajder (2015) seek for the prominent arguments in online debates. Here, we are not interested in the quality of single arguments but rather in the quality of a complete argumentation, namely, the argumentation found in a persuasive essay.

We target quality dimensions of persuasive essays that are directly related to argumentation: organization (Persing et al., 2010), thesis clarity (Persing and Ng, 2013), prompt adherence (Persing and Ng, 2014), and argument strength (Persing and Ng, 2015). In all four publications, sophisticated features are engineered to address a respective essay scoring task. The argument strength approach adopts ideas from the approach of Stab and Gurevych (2014b), but it finds structure heuristically only and, thus, does not perform argument mining. In the paper at hand, we fill this gap, i.e., we exploit the output of an argument mining approach trained on ground-truth data to assess the four quality dimensions.

In general, numerous approaches exist that assess essay quality. Classical essay scoring often focuses on grammar, vocabulary, and similar (Dikli, 2006), partly employing structural features like discourse markers (Burstein et al., 1998). In contrast, Song et al. (2014) study whether essays comply with critical

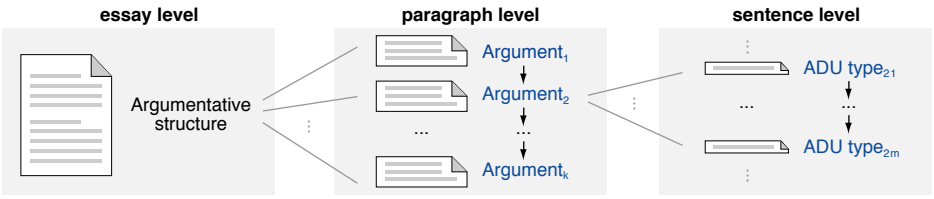

Figure 2: Application-oriented model of the argumentative structure of essays. Each paragraph is seen as an argument, defined as a sequence of sentence-level ADU types $\in \{$*Thesis, Conclusion, Premise, None*$\}$.

questions of an applied argumentation scheme. On manual annotations, they find correlations between an essay's score and the number of answered questions. Closer to our work, Ong et al. (2014) analyze argumentative discourse units found with a simple heuristic algorithm. And Ghosh et al. (2016) even derive features from argument mining, although they hardly exploit structure. Either way, all these approaches assign overall essay scores only, leaving unclear to what extent argumentation quality is captured.

3 Mining Argumentative Structure

This paper does *not* aim at new approaches to argument mining. Still, the effectiveness of mining as well as the underlying argumentation model directly affect the analysis of argumentative structure. Therefore, we summarize our mining approach in the following.[1]

3.1 An Application-Oriented Model of Argumentative Structure

We focus on the argumentative structures of persuasive student essays. Such an essay states and justifies a thesis on some topic that is introduced by a given prompt. To capture an essay's structure, we build on the work of Stab and Gurevych (2014a) who presented both an argumentation model for persuasive essays and an annotated corpus. By training a mining approach on this corpus, we expect to minimize the usual out-of-domain effectiveness drop (Blitzer et al., 2008), when using the approach on other essays.

Stab and Gurevych (2014a) distinguish four types of argumentative discourse units (called ADUs from here on) within essays: *Thesis*, *Conclusion*, *Premise*, and *None*.[2] The authors define an ADU loosely as a statement covering an entire sentence or less. Each conclusion in an essay supports or attacks a thesis, and each premise supports or attacks a thesis, conclusion, or other premise. Implicitly, these relations specify the essay's arguments. In their corpus, less than 15% of all relations are attacks.

For our purposes, we simplify the model of Stab and Gurevych (2014b) in two respects: (1) We define each sentence in an essay to correspond to exactly one ADU. Thereby, we avoid the need to segment essays into ADUs.[3] (2) We define each paragraph in an essay to correspond to exactly one argument. Thereby, we avoid the need to identify relations between ADUs. As a result, we represent the argumentative structure of an essay as a sequence of arguments and each argument as a sequence of ADU types. Figure 2 sketches this application-oriented model.

The justification for our simplification is twofold: (1) We aim to capture argumentative structure only on an abstraction level that allows assessing argumentation quality. Abstraction reduces the search space of argument structures to explore, which benefits pattern recognition, but it also takes away information. While the right level is unknown, we hypothesize that students largely organize essays sequentially. This is in line with our previous research (Wachsmuth et al., 2015). (2) For successful pattern recognition, we need to mine argumentative structure effectively. Therefore, we omit potentially helpful structure such as attack relations, as all available data seems insufficient for reliably training respective approaches.

3.2 Approach

For tokenization, sentence splitting, and paragraph splitting, we apply our own algorithms from previous work (Wachsmuth, 2015), while we use the TreeTagger for part-of-speech tagging (Schmid, 1995). Given the sentences and paragraphs of an essay, our model then requires only to classify the ADU type of

[1] The source code for reproducing all experiments from Sections 3 to 5 can be found here: http://www.arguana.com/software

[2] Stab and Gurevych (2014a) use other names for the ADU types than we do, such as *Major claim* instead of *Thesis*.

[3] The approach proposed by Stab and Gurevych (2014b) also does not deal with the segmentation of an essay into ADUs, but merely because it classifies ADUs simply based on the ground-truth segmentation.

ADU Type	Training	Test	Total	AAE Total
Thesis	72	18	90	90
Conclusion	325	93	418	429
Premise	652	181	833	1033
None	185	53	238	327
All types	1234	345	1579	1879

Table 1: Distribution of ADU type annotations in the modified dataset. Notice that the difference to the distribution in the original Argument Annotated Essays (AAE) corpus is moderate only.

#	Feature Type	Accuracy	F_1-score
1	Prompt similarity	44.9	41.8
2	Token n-grams	47.8	48.0
3	POS n-grams	41.2	43.5
4	General Inquirer classes	42.3	44.5
5	1st token n-grams	33.6	35.0
6	Sentence position	64.9	66.9
1–6	**Complete feature set**	74.5	**74.5**
	Majority baseline	52.5	36.1
	Stab and Gurevych (2014b)	**77.3**	72.6

Table 2: Effectiveness of our features in classifying ADU types compared to (Stab and Gurevych, 2014b).

each sentence. As Stab and Gurevych (2014b), we tackle this 4-class classification task with supervised machine learning. We employ six feature types that capture the content, style, and position of a sentence:[4]

Prompt Similarity The cosine, Euclidean, Manhattan, and Jaccard similarity of the sentence to the prompt of the given essay, once for all words and once for all non-function words.

Token n-Grams The frequency of each token 1- to 3-gram occurring in $\geq 1\%$ of the training sentences.

POS n-Grams The frequency of each part-of-speech 1- to 3-gram occurring in $\geq 5\%$ of these sentences.

General Inquirer Classes The frequency of each word class specified by the General Inquirer.[5]

1st Token n-Grams Indicators whether the first token 1-, 2-, and 3-gram of the sentence match those 1-, 2-, and 3-grams that are first in $\geq 0.5\%$ of all training ADUs.

Sentence Position Indicators whether a sentence is the first, second, or last within a paragraph and what its relative position is. The same for the sentence and the covering paragraph within the complete essay.

3.3 Experimental Set-up

We evaluated our approach to classify all ADU types in a persuasive essay based on the following set-up:

Data As indicated, we processed the Argument Annotated Essays (AAE) corpus of Stab and Gurevych (2014a), containing 90 persuasive student essays (72 for training, 18 for testing). In each essay, all theses, conclusions, and premises are annotated as ADUs of the respective types. Since we do not tackle ADU segmentation, we enlarged the annotations to span the whole covering sentence. If a sentence contained more than one ADU, we favored rarer classes to benefit training, i.e., we preferred *Thesis* over *Conclusion* over *Premise*. All unannotated sentences from an essay's body were assigned the type *None*. Unlike Stab and Gurevych (2014b), we ignored the titles of the 90 essays as *None* instances; classifying a title based on its position is trivial, but it causes errors on essays without titles. Table 1 compares the numbers of annotations in our modified dataset to those of the original AAE corpus. Besides the ignored titles, the two resources differ considerably only in the number of premises.

Experiments For supervised learning, we used the default configuration of the SMO classifier in Weka 3.7 (Hall et al., 2009). We turned off its feature normalization, though, because we generally normalize all our feature values to the range $[0, 1]$. On the training set of the derived dataset, we trained one classifier for each single feature type and for the complete feature set. We did not optimize any hyperparameters but simply measured the accuracy and weighted average F_1-score of the default SMO on the test set.

Comparison As a rough estimate, we compare the results of our approach to those of Stab and Gurevych (2014b) on the AAE corpus. While the comparability is only limited due to the slightly modified corpus, we do not primarily aim to outperform existing mining approaches but rather to imitate them. In order to ease the global interpretation of our results, we also report on the *majority baseline*.

3.4 Results

Table 2 presents the classification effectiveness of each evaluated feature type. The sentence position features dominate all other types with an accuracy of 64.9 and an F_1-score of 66.9. Still, the others add

[4]The strongest type in (Stab and Gurevych, 2014b) uses the length of an ADU as well as the tokens in its covering sentence. As we classify complete sentences, these features help less here.

[5]For more information on the General Inquirer classes, see http://www.wjh.harvard.edu/∼inquirer/.

Paragraph	Premise: *Secondly, most violent crimes are related to the abuse of guns, especially in some countries where guns are available for people.* Conclusion: *Eventually, guns will create a violent society if the trend continues.* Premise: *Take an example, in American, young adults and even juveniles can get access to guns, which leads to the tragedies of school gun shooting.* Premise: *What is worse, some terrorists are able to possess more advanced weapons than the police, which makes citizens always live in danger.*
ADU flow	(1x Premise, 1x Conclusion, 2x Premise)
ADU change flow	(Premise, Conclusion, Premise)

Figure 3: The ADU flow and the ADU change flow for one paragraph of the AAE corpus (see Section 3).

to the effectiveness of the complete feature set. The complete feature set performs a little worse than Stab and Gurevych (2014b) in terms of accuracy (74.5 vs. 77.3) but better in terms of F_1-score (74.5 vs. 72.6). Thus, we conclude that our mining approach is at eye level with (Stab and Gurevych, 2014b). Moreover, our results appear reasonable within a 4-class classification task. We will see whether they suffice to recognize discriminative argumentative structures and to leverage them for quality assessment.

4 Analyzing Argumentative Structure

This section analyzes the output of our mining approach to find statistically reliable patterns in the argumentative structure of persuasive essays. From these, novel features for machine learning are derived.

4.1 Statistically Reliable Patterns of Argumentative Structure

A persuasive essay is meant to compose a set of arguments in favor of or against a thesis, each combining a set of premises with a conclusion (Stab and Gurevych, 2014a). Such a tree-like structure allows for much variance, rendering a reliable pattern recognition hard. Above, we have hypothesized that essays largely argue sequentially. Given the model from Section 3, we hence restrict our view to the sequences of types of argumentative discourse units (ADUs) in essays. In accordance with our work on sentiment flows from (Wachsmuth et al., 2014b), we look at two kinds of patterns, both exemplified in Figure 3:

ADU Flow The sequence of all ADU types within one paragraph on an essay.

ADU Change Flow The sequence of all different ADU types within one paragraph on an essay.

4.2 Experimental Set-up

To get reliable insights into the structure of persuasive essays, we performed a straightforward analysis:

Data We took the International Corpus of Learner English (ICLE, version 2), containing 6085 English essays from students of 16 mother tongues (Granger et al., 2009). On average, an ICLE essay spans 7.6 paragraphs (standard deviation $\pm$ 5.2) and 33.8 sentences ($\pm$ 16.5) according to our preprocessing.

Experiments We applied the mining approach from Section 3 to all ICLE essays. Then, we computed the relative frequencies of all ADU flows and ADU change flows. In addition, we tested how much these frequencies differ within an essay's first and last paragraph.

4.3 Results

The top part of Table 3 lists the ten most frequent of the 2593 ADU flows found in the ICLE corpus. They cover about half of all paragraphs. The first two ADU flows consist of conclusions only, whereas the others show "real" argumentative structure. After a conclusion, two premises follow most often (5.4%). Still, the number of premises varies, bringing up the question whether a particular number benefits argumentation quality. Patterns such as *(1x Conclusion, 2x Premise, 1x Conclusion)* may refer to restated conclusions but also to paragraphs that combine two arguments.

Overall, we see that all top ten ADU flows begin with a conclusion, i.e., our analysis reveals that students tend to (or are taught to) first state a claim and then argue for it. This is also supported by the top ten ADU change flows in Table 4: Every fourth paragraph matches the pattern *(Conclusion, Premise)*, while only 2.9% order all premises first. Similarly, *None* serves for beginning a paragraph, while theses rather appear at the end. In total, the abstraction of ADU change flows seems to capture much diversity of arguments in persuasive essays: Together, the top ten represent 87.4% of all ICLE paragraphs, and all ADU types occur in at least one combination. Still, we found 319 different ADU change flows.

Both Table 3 and 4 highlight the special roles of the first and last paragraph of an essay, which clearly deviate from the average: The first is mostly made up of *None* and *Thesis*, underlining its introductory nature. In contrast, the last often ends with a conclusion—making the argumentation's final point.

¶'s	#	ADU Flow	Frequency
all	1	(1x Conclusion)	14.5%
	2	(2x Conclusion)	7.1%
	3	(1x Conclusion, 2x Premise)	5.4%
	4	(1x Conclusion, 1x Premise)	4.9%
	5	(1x Conclusion, 3x Premise)	4.2%
	6	(1x Conclusion, 1x Premise, 1x Conclusion)	4.2%
	7	(1x Conclusion, 2x Premise, 1x Conclusion)	3.4%
	8	(1x Conclusion, 4x Premise)	3.0%
	9	(1x Conclusion, 3x Premise, 1x Conclusion)	2.3%
	10	(1x Conclusion, 5x Premise)	2.0%
1st	1	(2x None)	9.7%
	2	(3x None)	8.6%
	3	(2x None, 1x Thesis)	6.2%
	4	(4x None)	6.2%
	5	(3x None, 1x Thesis)	5.7%
last	1	(1x Conclusion)	16.9%
	2	(2x Conclusion)	12.6%
	3	(1x Conclusion, 1x Premise, 1x Conclusion)	8.2%
	4	(1x Conclusion, 2x Premise, 1x Conclusion)	5.7%
	5	(1x Conclusion, 3x Premise, 1x Conclusion)	3.4%

Table 3: The most frequent ADU flows in *all* ICLE paragraphs as well as in the *1st* and *last* paragraphs.

¶'s	#	ADU Change Flow	Frequency
all	1	(Conclusion, Premise)	25.1%
	2	(Conclusion)	22.4%
	3	(Conclusion, Premise, Conclusion)	17.0%
	4	(None)	5.8%
	5	(Premise)	4.3%
	6	(None, Thesis)	3.4%
	7	(Premise, Conclusion)	2.9%
	8	(None, Premise)	2.7%
	9	(Conclusion, Premise, Conclusion, Premise)	2.0%
	10	(None, Premise, Conclusion)	1.8%
1st	1	(None)	42.7%
	2	(None, Thesis)	25.9%
	3	(Thesis)	5.7%
	4	(None, Premise, None)	4.4%
	5	(None, Conclusion)	4.3%
last	1	(Conclusion)	31.6%
	2	(Conclusion, Premise, Conclusion)	27.2%
	3	(Conclusion, Premise)	13.1%
	4	(None, Premise, Conclusion)	4.4%
	5	(Premise, Conclusion)	2.7%

Table 4: The most frequent ADU change flows in the ICLE paragraphs (ignoring type repititions).

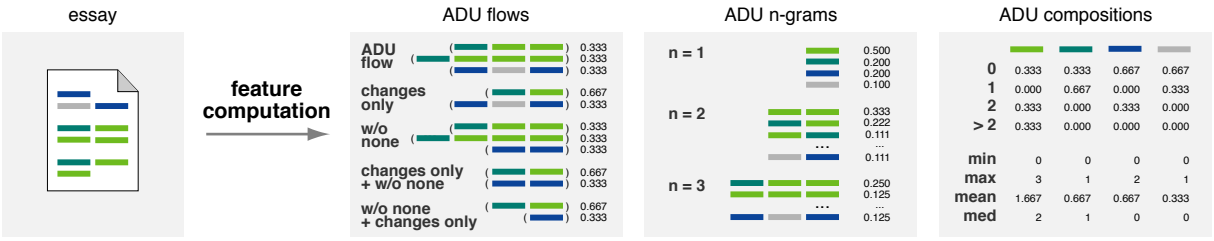

Figure 4: Sketch of the three feature types that we propose based on the output of argument mining.

4.4 Shallow Features for Statistical Significance

The found patterns suggest that persuasive student essays differ in the combination, ordering, and number of ADU types. For the assessment of argumentation quality, we capture these structural variations in the following three novel feature types. The types are kept shallow in order to benefit statistical significance:

ADU Flows The frequencies of all ADU flows in an essay. The hypothesis is that certain flows are favorable. We also examine two flow abstractions: (1) considering changes only, as above, and (2) ignoring the non-argumentative type *None*. Arranging 0 to 2 of these abstractions allows for five flow variations.

ADU n-Grams The frequencies of all ADU type n-grams in an essay for some $n \geq 1$. The hypothesis is that certain combinations of ADU types are favorable.

ADU Compositions The proportions of paragraphs in an essay with a particular number of occurrences of a particular ADU type as well as summary statistics about each type (such as the minimum or mean). The hypothesis is that certain numbers of certain ADU types are favorable.

Figure 4 illustrates the computation of feature values of each type for a sample essay with three paragraphs. The exact feature type configuration that we used in our experiments is specified in Section 5.

5 Assessing Argumentation Quality

Finally, we analyze the benefit of mining argumentative structure for assessing argumentation quality. In particular, we evaluate the three presented feature types in argumentation-related essay scoring.

5.1 Essay Scoring Tasks

We consider four essay scoring tasks that were introduced in successive papers, each of which capturing a particular dimension of argumentation quality. These tasks can be summarized as follows:

Organization Score the quality of an essay's organization. A high score is assigned to essays, which introduce their topic, take and argue for a position on the topic, and conclude (Persing et al., 2010).

Thesis Clarity Score the clarity of the explanation of the thesis that an essay argues for. A high score is assigned to essays, which make their thesis easy to understand (Persing and Ng, 2013).

Prompt Adherence Score the adherence of an essay's content to the essay's prompt. A high score is assigned to essays that consistently remain on the topic of the prompt (Persing and Ng, 2014).

Argument Strength Score the strength of the argument that an essay makes for its thesis. A high score is assigned to essays, which would convince most readers of their thesis (Persing and Ng, 2015).

Our proposed feature types solely focus on the argumentative structure of an essay—as opposed to the essay's content or linguistic style. Accordingly, we hypothesize that the feature types are particularly successful in the organization task. To a minor extent, we expect that they also help for argument strength, because argument strength should emerge from all aspects of an essay. In contrast, the scoring of thesis clarity and prompt adherence rather seems to require an analysis of content and style respectively.

5.2 Approach

Analogue to the authors of the four mentioned papers, we tackle essay scoring with supervised regression. For this purpose, we consider our proposed feature types as well as several baseline features:

ADU Features (a) In terms of the feature types from Section 4, we rely on the following configurations:

- a_1 *ADU flows*. The frequency of each ADU flow that occurs in $\geq 1\%$ of all training essays. All five flow variations described in Section 4 are taken into account.
- a_2 *ADU n-grams*. The frequency of each ADU 1-, 2-, and 3-gram that occurs in $\geq 5\%$ of all training essays. + Indicators that capture the first and the last ADU 1-, 2-, and 3-gram.
- a_3 *ADU compositions*. The percentages of paragraphs with $\{0 \mid 1 \mid 2 \mid >2\}$ occurrences of the type $\{$*Thesis* | *Conclusion* | *Premise* | *None*$\}$. + The {minimum | maximum | mean | median} of each of these ADU types per paragraph. + The percentage of each type in the first and in the last paragraph.

Flow Features (b) Persing et al. (2010) aligned sequences of four paragraph discourse functions: *Body* (own argument), *Rebuttal*, *Introduction*, and *Conclusion*. Since we cannot access their original approach, we approximate it—and also add further strong structure-oriented baseline approaches: In (Wachsmuth et al., 2014a) and (Wachsmuth et al., 2015), we captured the overall structure of a review by comparing the review's sentiment flow to a set of common flow patterns and flow abstractions. Both the patterns and the abstractions were found in a training set before. To model the paragraph-level argumentative structure of persuasive essays, we adapt these patterns and abstractions in the following features:

- b_1 *Function flows*. All flow features defined in (Wachsmuth et al., 2014a) and (Wachsmuth et al., 2015) based on paragraph discourse functions. Functions are found with the heuristic algorithm of Persing et al. (2010). *Body* is mapped to 1.0, *Rebuttal* to 0.0, and the remaining two functions to 0.5, in order to allow for numerical comparison between the flows.
- b_2 *Sentiment flows*. All flow features based on paragraph-level sentiment. A paragraph is assigned the numerical sentiment value 1.0 (0.0), if it contains a positive (negative) but no negative (positive) sentence, otherwise 0.5. We find sentence sentiment with the algorithm of Socher et al. (2013).
- b_3 *Relation flows*. All flow features based on sentence-level discourse relations. Ten relation types from (Mann and Thompson, 1988) are found with our rule-based algorithm (Wachsmuth et al., 2014a). For lack of an adequate mapping, we compare relation flows based on nominal differences only.

Standard Features (c) In order to be able to assess the impact of argumentative structure, we compare all structure-oriented features to two standard types of content and style features:

- c_1 *Content*. The frequency of each token 1-, 2-, and 3-gram that occurs in $\geq 10\%$ of all training essays. + The minimum, maximum, and average prompt similarity (see Section 3.2) over all sentences.
- c_2 *POS n-grams*. The frequency of each part-of-speech 1-, 2-, and 3-gram that occurs in $\geq 10\%$, $\geq 20\%$, and $\geq 40\%$ of all training essays respectively.

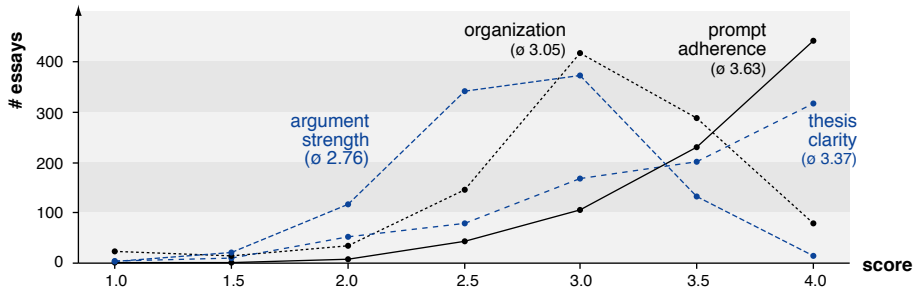

Figure 5: Distribution of essays over the possible scores from [1.0, 4.0] in the datasets of the four tasks.

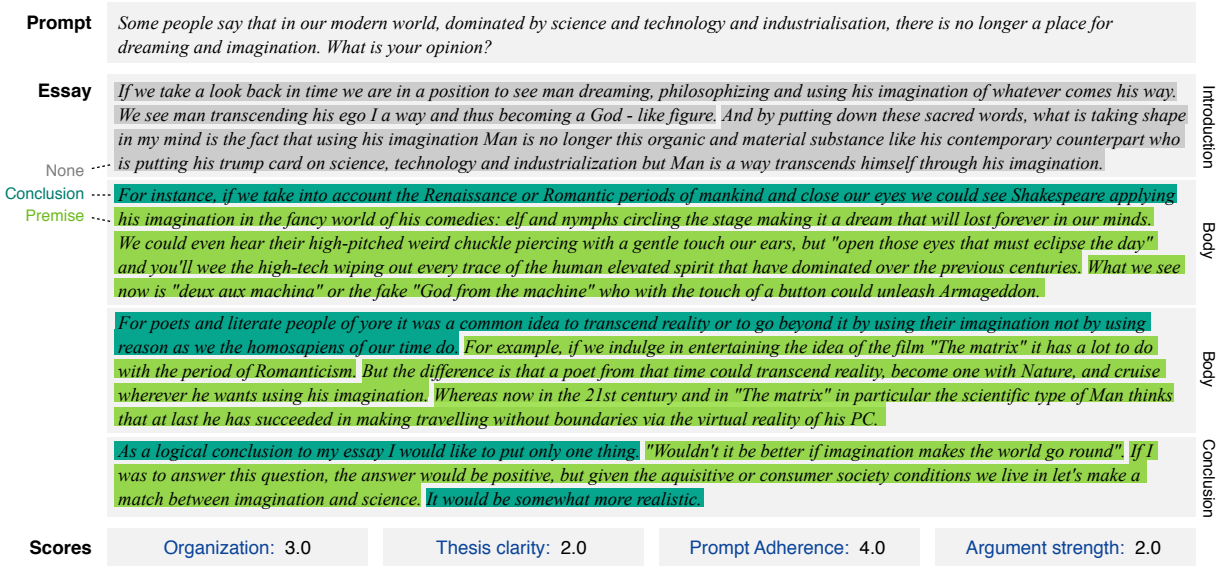

Figure 6: One essay from the four datasets together with its manually assigned scores as well as the ADU types (colored background) and discourse functions (vertical) automatically annotated by our algorithms.

5.3 Experimental Set-up

For direct comparison, we replicated the original experimental set-up of the authors of the aforementioned papers on the datasets they provide for the four essay scoring tasks:

Data For each task, one distinct subset of the ICLE corpus (see Section 4) is manually annotated with half-point scores between 1.0 (worst) and 4.0 (best). These datasets cover 1003 (organization), 830 (thesis clarity, prompt adherence), and 1000 essays (argument strength) respectively. For a rough overview, Figure 5 plots the numbers of scores in each dataset, indicating that only the organization and argument strength scores are Gaussian-like distributed. Exact numbers are found in the original papers. Figure 6 shows one essay included in all datasets with its scores and the annotations created by our algorithms.

Experiments We used linear ε-SVR support vector machine regression from LibSVM in Weka 3.7 (Hall et al., 2009; Chang and Lin, 2011).[6] Each dataset has five predefined folds. As in the original set-up, we performed cross-validation on these folds, training one LibSVM for each feature type and for different type combinations. Accordingly, we then also measured the mean absolute error (MAE) and the mean squared error (MSE) of regression.[7] Different from the original set-up, we omitted a real optimization of the LibSVM cost hyperparameter, but we simply set it permanently to 0.1 after a few initial tests.

Comparison We compare the proposed feature types to the described baseline features and to two general baselines: (**d**) The *average baseline*, which assigns the mean score of the training essays to all test essays. Although trivial, **d** is quite strong under given the score distributions in Figure 5. (**e**) The lowest MAE and MSE values reported by the authors of the four tasks, called *Persing et al. best* below. To our knowledge, these results have not been beaten so far and, thus, define the state of the art until now.

[6] Persing and Ng (2014; 2015) relied on LibSVM, too. In the other two papers, SVM^{light} was used (Joachims, 1999).

[7] From the practical viewpoint of applying automatic essay scoring in MOOCs or similar, the most important requirement is to avoid outliers (in terms of utterly wrong scores) as far as possible. In this regard, the MSE is the more meaningful measure.

#	Feature Type	Organization MAE	Organization MSE	Thesis Clarity MAE	Thesis Clarity MSE	Prompt Adherence MAE	Prompt Adherence MSE	Argument Strength MAE	Argument Strength MSE
a_1	ADU flows	0.367 ±.022	0.234 ±.017	0.530 ±.030	0.461 ±.061	0.373 ±.033	0.247 ±.056	0.399 ±.010	**0.242** ±.018
a_2	ADU n-grams	0.369 ±.024	0.225 ±.031	0.530 ±.032	0.466 ±.065	0.372 ±.031	0.265 ±.053	0.398 ±.012	**0.243** ±.017
a_3	ADU compositions	0.347 ±.026	0.194 ±.020	0.529 ±.034	0.457 ±.062	0.365 ±.032	0.239 ±.046	**0.390** ±.015	**0.239** ±.023
a	**ADU features**	0.336 ±.022	0.184 ±.020	0.537 ±.031	0.470 ±.056	0.368 ±.029	0.241 ±.044	0.392 ±.016	**0.242** ±.023
b_1	Function flows	0.368 ±.025	0.220 ±.037	0.541 ±.032	0.478 ±.063	0.370 ±.029	0.255 ±.042	0.403 ±.006	0.251 ±.011
b_2	Sentiment flows	0.369 ±.014	0.228 ±.022	0.536 ±.031	0.481 ±.064	0.372 ±.035	0.257 ±.053	0.410 ±.009	0.259 ±.013
b_3	Relation flows	0.426 ±.013	0.351 ±.033	0.541 ±.052	0.475 ±.074	0.368 ±.027	0.255 ±.048	0.408 ±.012	0.260 ±.014
b	**Flow features**	0.355 ±.023	0.207 ±.026	0.559 ±.040	0.512 ±.059	0.377 ±.020	0.255 ±.038	0.415 ±.009	0.269 ±.013
c_1	Content	0.415 ±.013	0.336 ±.025	0.501 ±.033	0.425 ±.064	0.362 ±.026	0.231 ±.046	0.395 ±.011	**0.236** ±.018
c_2	POS n-grams	0.407 ±.013	0.326 ±.032	0.528 ±.036	0.461 ±.063	0.361 ±.034	0.231 ±.048	**0.387** ±.014	**0.233** ±.019
c	**Standard features**	0.408 ±.013	0.324 ±.027	0.505 ±.036	0.429 ±.065	0.357 ±.026	0.222 ±.040	**0.384** ±.015	**0.230** ±.020
	a + b_1	**0.321** ±.020	**0.171** ±.019	0.551 ±.036	0.494 ±.064	0.372 ±.025	0.245 ±.038	0.400 ±.011	0.250 ±.019
	a + b_2	0.328 ±.016	**0.174** ±.016	0.546 ±.024	0.495 ±.049	0.374 ±.026	0.244 ±.043	0.402 ±.013	0.253 ±.018
	a + b_3	0.341 ±.019	0.189 ±.018	0.544 ±.035	0.482 ±.067	0.367 ±.027	0.240 ±.043	0.405 ±.022	0.254 ±.022
	a + **b**	0.329 ±.018	0.179 ±.020	0.565 ±.026	0.484 ±.057	0.379 ±.021	0.250 ±.037	0.421 ±.016	0.275 ±.016
	a + b_1 + c_1	**0.315** ±.018	**0.168** ±.020	0.524 ±.031	0.456 ±.057	0.361 ±.021	0.227 ±.033	**0.391** ±.010	**0.242** ±.019
	a + b_1 + c_2	**0.315** ±.017	**0.167** ±.017	0.546 ±.031	0.492 ±.056	0.362 ±.025	0.229 ±.033	**0.389** ±.011	**0.241** ±.020
*	**a** + b_1 + **c**	**0.314** ±.018	**0.167** ±.018	0.520 ±.032	0.450 ±.056	0.362 ±.021	0.226 ±.030	**0.387** ±.012	**0.238** ±.021
	a + b_2 + c_1	**0.320** ±.016	**0.167** ±.016	0.520 ±.023	0.457 ±.049	0.367 ±.024	0.230 ±.039	0.395 ±.013	0.247 ±.019
*	**a** + b_2 + c_2	**0.316** ±.015	**0.164** ±.013	0.548 ±.020	0.496 ±.046	0.364 ±.029	0.232 ±.041	0.393 ±.012	0.246 ±.019
	a + b_2 + **c**	**0.320** ±.014	**0.167** ±.014	0.520 ±.025	0.456 ±.051	0.363 ±.026	0.228 ±.037	0.392 ±.011	**0.243** ±.020
	a + b_3 + c_1	0.333 ±.020	0.182 ±.017	0.520 ±.031	0.443 ±.060	0.360 ±.023	0.228 ±.037	0.397 ±.014	0.247 ±.023
	a + b_3 + c_2	0.326 ±.018	0.176 ±.014	0.537 ±.030	0.477 ±.057	0.363 ±.027	0.232 ±.039	0.393 ±.014	0.244 ±.024
	a + b_3 + **c**	0.328 ±.017	0.177 ±.014	0.515 ±.032	0.441 ±.054	0.359 ±.023	0.226 ±.037	0.390 ±.014	**0.243** ±.024
	a + **b** + c_1	**0.321** ±.021	**0.172** ±.018	0.543 ±.029	0.484 ±.055	0.376 ±.018	0.242 ±.033	0.413 ±.008	0.265 ±.014
	a + **b** + c_2	**0.315** ±.019	**0.169** ±.016	0.557 ±.024	0.512 ±.049	0.375 ±.021	0.243 ±.033	0.411 ±.008	0.262 ±.014
	a + **b** + **c**	**0.315** ±.018	**0.169** ±.015	0.543 ±.029	0.486 ±.054	0.375 ±.017	0.241 ±.032	0.409 ±.008	0.259 ±.014
	a_3 + c_1	0.335 ±.022	0.182 ±.018	0.505 ±.036	0.431 ±.061	0.356 ±.029	0.221 ±.043	**0.383** ±.016	**0.230** ±.024
**	a_3 + c_2	0.329 ±.018	0.177 ±.013	0.529 ±.033	0.463 ±.062	0.354 ±.033	0.221 ±.042	**0.380** ±.016	**0.229** ±.024
**	a_3 + **c**	0.330 ±.017	0.178 ±.014	0.508 ±.036	0.435 ±.061	0.352 ±.027	0.216 ±.038	**0.378** ±.017	**0.226** ±.025
	b + c_1	0.344 ±.024	0.196 ±.026	0.531 ±.014	0.464 ±.066	0.372 ±.015	0.242 ±.015	0.406 ±.012	0.257 ±.015
	b + c_2	0.336 ±.018	0.189 ±.022	0.552 ±.013	0.450 ±.056	0.369 ±.022	0.242 ±.017	0.404 ±.015	0.255 ±.015
	b + **c**	0.336 ±.018	0.189 ±.021	0.531 ±.015	0.465 ±.064	0.368 ±.017	0.237 ±.016	0.400 ±.015	0.250 ±.015
d	**Average baseline**	0.425 ±.016	0.349 ±.030	0.545 ±.036	0.469 ±.084	0.370 ±.038	0.291 ±.055	0.407 ±.014	0.266 ±.018
e	**Persing et al. best**	0.323	0.175	0.483	0.369	0.348	0.197	0.392	0.244

Table 5: The mean average error (MAE) and the mean squared error (MSE) ± their standard deviations for each evaluated feature type and type combination in the four essay scoring tasks. All values marked in bold outperform the former state of the art (called *Persing et al. best* here). The most significant results for organization and argument strength each are marked with * and ** respectively.

5.4 Results

Table 5 lists the two mean regression errors and their standard deviations for each single feature type in isolation and for a selection of feature type combinations, averaged over the respective five folds for each of the four evaluated essay scoring tasks.[8]

With respect to the single feature types in the upper part, we see that the ADU compositions (a_3) consistently perform best among all structure-oriented features (a_1–a_3, b_1–b_3) in all four tasks. The ADU flows (a_1) and ADU n-grams (a_2) behave comparable to the function flows (b_1) and sentiment flows (b_2). The relation flows (b_3) compete with most others only in scoring prompt adherence, possibly suggesting that the impact of discourse structure on argumentation quality is limited.[9]

Matching our hypothesis from Section 5.1, the structure-oriented features clearly outperform the standard features (**c**) in scoring *organization*, as shown in the left part of Table 5. In fact, c_1 and c_2 hardly improve over the average baseline (**d**). In isolation, the complete ADU feature set (**a**) produces the lowest errors; an MAE of 0.336 and an MSE of 0.184. Since the quality of the organization of an essay naturally depends on the essay's argumentative structure, this result underpins the adequacy of our features. Moreover, **a** outperforms the best values we measured without using ADU features (those of **b** + **c**). Combined with function flows (**a** + b_1) or with sentiment flows (**a** + b_2), the ADU features already beat Persing et

[8] We did not perform explicit feature ablation tests, because they would provide limited insights only: As we did not optimize the LibSVM cost hyperparameter, leaving out one feature type does not necessarily lead to an increase of the regression errors.

[9] One reason for the low impact of b_3 lies in the varying number of sentences of essays (see Section 4), which impairs flow pattern recognition. A detailed analysis of the use of discourse relations for quality assessment is out of the scope of this paper.

al. best (**e**). The smallest MSE is achieved by the ADU features with sentiment flows and POS n-grams ($\mathbf{a} + b_2 + c_2$). According to a one-sided student t-test, the value 0.164 is significant at $p < 0.1$.

As expected, all structure-oriented features fail in case of *thesis clarity*, being only slightly better than the average baseline (**d**) if at all. The lowest errors (MAE 0.501, MAE 0.425) are observed for the content features (c_1). Still, c_1 cannot compete with **e**—the respective approach of Persing and Ng (2013) employs keyword features that were manually derived from the prompts of all essays.

For *prompt adherence*, at least the errors produced by the ADU compositions (a_3) are close to those of c_1 and c_2, which is why we additionally tested a_3 in combination with the standard feature types. As shown in the bottom part of Table 5, $a_3 + \mathbf{c}$ performs best with an MAE of 0.352 and an MSE of 0.216. These values are not significantly worse than the state of the art (**e**).

$a_3 + \mathbf{c}$ also minimizes the errors in scoring *argument strength*. Both the MAE of 0.378 and the MSE of 0.226 are significantly better than Persing et al. best (**e**) at $p < 0.1$. Again, this observation supports our hypothesis: The strength of an essay's argumentation will hardly ever be independent from the essay's content, but it still benefits from a good argumentative structure. Interestingly, even a_3 alone improves over **e** with an MAE of 0.390 (vs. 0.392) and an MSE of 0.239 (vs. 0.244).

We conclude that our approach denotes the new state of the art for two essay scoring tasks. Under the assumption that the manual score annotations in the processed datasets are adequate, our hypothesis that the benefit of argument mining is high for scoring an essay's organization turns out true. Compared to the findings of Persing et al. (2010), the obtained results thereby reveal that organization is not only about the ordering of discourse functions, but also about argumentative structure. In particular, the novel features that we proposed capture only such structural aspects. Accordingly, their impact is low for thesis clarity and also only fair for prompt adherence, underlining that these tasks are rather related to the content and style of an argumentation. In contrast, argument strength brings together structure and content, and this is indeed reflected by the moderate but significant benefit of our structure-oriented features there.

6 Conclusion

Although argument mining has become a hot topic, the question of what practical benefits it provides for applications has hardly been examined yet. In the paper at hand, we have approached this question for a specific but important task, namely, we have used argument mining to assess argumentation quality. Our results for persuasive student essays underpin the benefit of argument mining, revealing that the mined argumentative structure is particularly helpful for structure-related quality dimensions: Without putting emphasis on the content of arguments, we have improved the state of the art in scoring an essay's organization and even in scoring its argument strength. Our best-performing features capture the composition of types of units in arguments (such as premises and conclusions).

Similar to existing approaches, the mining algorithm we trained and applied in this paper misclassifies about one out of four units. So far, we could not analyze the impact of mining errors on the effectiveness of our essay scoring approaches, since ground-truth data is needed before that brings together argumentative structure and argumentation quality. A question that remains open in this regard is what model of argumentative structure proves most suitable. As adequate training data is still limited, we have modeled shallow unit types only, but we expect that considering attack and support relations, evidence types, or argumentation schemes will prove useful for quality assessment.

Naturally, other quality dimensions of argumentation will depend more on content, so our analysis of argumentative structure does not solve the assessment of argumentation quality in general. Also, essay structure is quite conventionalized, i.e., a transfer of our findings to other argumentative text genres requires further investigation. We plan to continue our research in this regard based on our new corpus for the analysis of argumentation strategies in news editorials (Al-Khatib et al., 2016).

In practice, our approach in its given form most notably contributes to educational applications that analyze argumentative texts, such as automatic grading and writing support systems. These systems need not only mine argumentative structure, but also evaluate the mined structure.[10] To support argumentation quality, another step is then to synthesize suggestions for improvements. We leave this to future work.

[10] A demo application based on our presented approaches is found at: http://webis16.medien.uni-weimar.de/essay-scoring

References

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A News Editorial Corpus for Mining Argumentation Strategies. In *Proceedings of the 26th International Conference on Computational Linguistics*.

J. Anthony Blair. 2012. Relevance, Acceptability and Sufficiency Today. In *Groundwork in the Theory of Argumentation*, pages 87–100.

John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jenn Wortman. 2008. Learning Bounds for Domain Adaptation. In *Advances in Neural Information Processing Systems 21*. MIT Press.

Filip Boltužić and Jan Šnajder. 2015. Identifying Prominent Arguments in Online Debates Using Semantic Textual Similarity. In *Proceedings of the Second Workshop on Argumentation Mining*, pages 110–115.

Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, and Martin Chodorow. 1998. Enriching Automated Essay Scoring Using Discourse Marking. In *Discourse Relations and Discourse Markers*, pages 15–21.

Elena Cabrio and Serena Villata. 2012. Combining Textual Entailment and Argumentation Theory for Supporting Online Debates Interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers*, pages 208–212.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

Danish Contractor, Yufan Guo, and Anna Korhonen. 2012. Using Argumentative Zones for Extractive Summarization of Scientific Articles. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 663–678.

Semire Dikli. 2006. An Overview of Automated Scoring of Essays. *Journal of Technology, Learning, and Assessment*, 5(1).

Phan Minh Dung. 1995. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence*, 77(2):321–357.

Adam Robert Faulkner. 2014. *Automated Classification of Argument Stance in Student Essays: A Linguistically Motivated Approach with an Application for Supporting Argument Summarization*. Dissertation, City University of New York.

Vanessa Wei Feng, Ziheng Lin, and Graeme Hirst. 2014. The Impact of Deep Hierarchical Discourse Structures in the Evaluation of Text Coherence. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, pages 940–949.

James B. Freeman. 2011. *Argument Structure: Representation and Theory*. Springer.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. 2016. Coarse-grained Argumentation Features for Scoring Persuasive Essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 549–554.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. International Corpus of Learner English (Version 2).

Ivan Habernal and Iryna Gurevych. 2015. Exploiting Debate Portals for Semi-Supervised Argumentation Mining in User-Generated Web Discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2127–2137.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18.

Thorsten Joachims. 1999. Advances in Kernel Methods. chapter Making Large-scale Support Vector Machine Learning Practical, pages 169–184. MIT Press, Cambridge, MA, USA.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, 8(3):243–281.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation Mining. *Artificial Intelligence and Law*, 19(1):1–22.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic Detection of Arguments in Legal Texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, pages 225–230.

Nathan Ong, Diane Litman, and Alexandra Brusilovsky. 2014. Ontology-Based Argument Mining and Automatic Essay Scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 24–28.

Andreas Peldszus and Manfred Stede. 2013. From Argument Diagrams to Argumentation Mining in Texts: A Survey. *International Journal of Cognitive Informatics and Natural Intelligence*, 7(1):1–31.

Andreas Peldszus and Manfred Stede. 2015. Joint Prediction in MST-style Discourse Parsing for Argumentation Mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948.

Isaac Persing and Vincent Ng. 2013. Modeling Thesis Clarity in Student Essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics - Volume 1: Long Papers*, pages 260–269.

Isaac Persing and Vincent Ng. 2014. Modeling Prompt Adherence in Student Essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics - Volume 1: Long Papers*, pages 1534–1543.

Isaac Persing and Vincent Ng. 2015. Modeling Argument Strength in Student Essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 543–552.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling Organization in Student Essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239.

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.

Helmut Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 69–78.

Christian Stab and Iryna Gurevych. 2014a. Annotating Argument Components and Relations in Persuasive Essays. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510.

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*, pages 46–56.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards Discipline-independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502.

Stephen E. Toulmin. 1958. *The Uses of Argument*. Cambridge University Press.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, and Gregor Engels. 2014a. Modeling Review Argumentation for Robust Sentiment Analysis. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, pages 553–564.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014b. A Review Corpus for Argumentation Analysis. In *Proceedings of the 15th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 115–127.

Henning Wachsmuth, Johannes Kiesel, and Benno Stein. 2015. Sentiment Flow – A General Model of Web Review Argumentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 601–611, Lisbon, Portugal.

Henning Wachsmuth. 2015. *Text Analysis Pipelines—Towards Ad-hoc Large-scale Text Mining*, volume 9383 of *Lecture Notes in Computer Science*. Springer.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.