Don't Drop the Topic - The Role of the Prompt in Argument Identification in Student Writing

Yuning Ding and **Marie Bexte** and **Andrea Horbach**
Research Cluster D^2L^2 - Digitalization, Diversity
and Lifelong Learning. Consequences for Higher Education.
FernUniversität in Hagen, Germany

Abstract

In this paper, we explore the role of topic information in student essays from an argument mining perspective. We cluster a recently released corpus through topic modeling into prompts and train argument identification models on various data settings. Results show that, given the same amount of training data, prompt-specific training performs better than cross-prompt training. However, the advantage can be overcome by introducing large amounts of cross-prompt training data.

1 Introduction

Argumentative essays are among the most common essay types students are assigned to write in higher education contexts (Wingate, 2012). In such an essay, students have to state and justify their opinion on a certain topic elicited by a specific writing prompt. In order to score argumentative essays and give formative feedback automatically, the automatic identification and classification of components in the argumentative structure is important (Scheuer et al., 2010). While their holistic scoring can be seen as one variant of **automatic essay scoring**, identifying the argumentative structure within an essay is a Natural Language Processing (NLP) task known as **argument mining**.

Argument mining is the automatic identification and extraction of the structure of inference and reasoning expressed as arguments presented in natural language (Lawrence and Reed, 2020). The recent Kaggle competition "Feedback Prize - Evaluating Student Writing"[1] can be seen as an argument mining task in an educational scenario, which called on participants to identify argumentative elements in English essays written by U.S. students. Figure 1 shows an example from the dataset for an essay where students have been asked to express their attitude towards driverless cars. Individual argumentative elements such as *Position*, *Evidence* or *Concluding Statement* are highlighted in the text.

The argument mining task is not restricted to a certain domain or topic. For example, previous work considered legal (Mochales and Ieven, 2009), political (Walker et al., 2012) or educational (Stab and Gurevych, 2017) data. However, it is an open question to what extent argument mining algorithms pick up on topical words indicative not for, e.g., a conclusion in general, but for a conclusion within a specific topic. The Feedback Prize data mentioned above with its large amount of annotated student essays on various topics offers an ideal opportunity for first steps towards closing this gap.

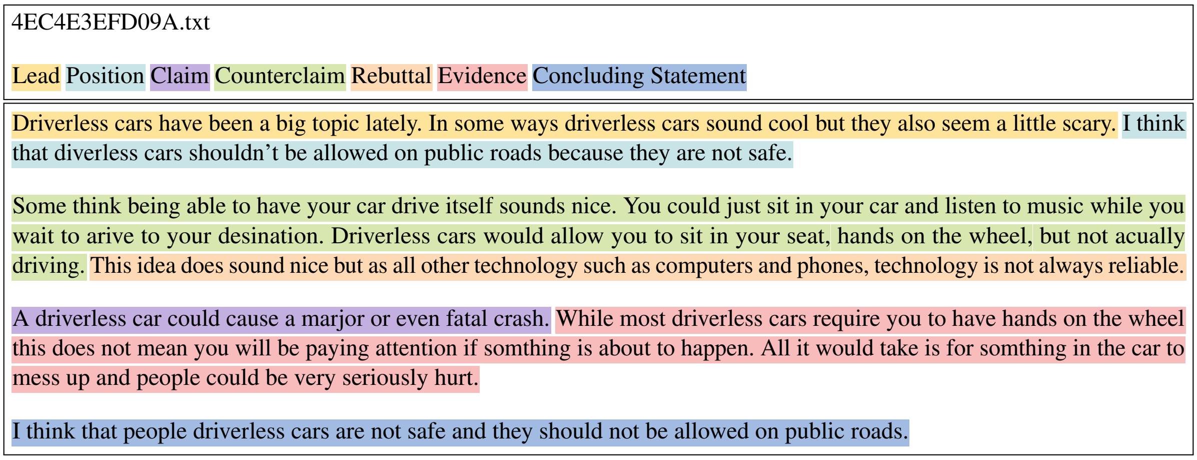
In the data, we notice that very similar sentences can receive different argumentative labels depending on the topic and the context of an essay. For example, the sentence *"exercise is really good for your health"* was annotated as a **claim** in an essay on the topic *"Limiting Car Usage"* while the sentence *"(. . .) running is good for your body"* was marked as **evidence** for the topic *"No Sports at Grade C"*. Such examples highlight the relevance of topic and context information for the argument mining task and give rise to research questions like:

- In how far is the task of argument mining prompt-dependent, i.e., how does prompt-specific vs. cross-prompt training affect classification performance?
- What kind of information is learned by an automatic argument classifier? Are algorithms more susceptible to prompt-specific words, or do they learn the general structure of an essay?

To address these questions, we present in this paper experimental studies to investigate the influence of the prompt in an educational argument identifying task using the example of the newly released Kaggle Feedback Prize dataset.

[1] https://www.kaggle.com/c/feedback-prize-2021

July 15, 2022 ©2022 Association for Computational Linguistics



4EC4E3EFD09A.txt

Lead Position Claim Counterclaim Rebuttal Evidence Concluding Statement

Driverless cars have been a big topic lately. In some ways driverless cars sound cool but they also seem a little scary. I think that diverless cars shouldn't be allowed on public roads because they are not safe.

Some think being able to have your car drive itself sounds nice. You could just sit in your car and listen to music while you wait to arive to your desination. Driverless cars would allow you to sit in your seat, hands on the wheel, but not acually driving. This idea does sound nice but as all other technology such as computers and phones, technology is not always reliable.

A driverless car could cause a marjor or even fatal crash. While most driverless cars require you to have hands on the wheel this does not mean you will be paying attention if somthing is about to happen. All it would take is for somthing in the car to mess up and people could be very seriously hurt.

I think that people driverless cars are not safe and they should not be allowed on public roads.

Figure 1: An example essay with different argumentative elements from the Kaggle competition "Feedback Prize - Evaluating Student Writing".

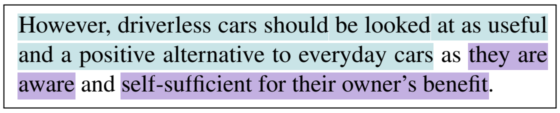
We find that argument mining benefits from within-prompt training data, but the same performance can be reached by using larger amounts of cross-prompt data. The argumentative elements *lead* and *conclusion* can be best identified because of their relatively fixed position within the essay. In an analysis of our models trained and tested with either topic or structure words masked, we find a tendency that within-prompt training benefits more from topic information while cross-prompt training rather picks up on structure words. We have made our experimental code, together with the automatic clustering results, publicly available at https://github.com/yuningDING/BEA-NAACL-2022-38.

2 Related Work

In the following, we discuss related work performing argument mining in the educational domain and work addressing the relevance of topic information.

Early work treated sentence boundaries as the natural separator of components in an essay. In such a scenario, the identification of argumentative elements boils down to a **sentence classification** task. For example, Burstein et al. (2003) classified sentences as *introductory material*, *position*, *main/supporting idea*, *conclusion*, *title* and *irrelevant* automatically, using features derived from Rhetorical Structure Theory trees and the occurrence of discourse markers. Ong et al. (2014) developed a rule-based algorithm to label each sentence in a student essay into one out of four types (*current study, hypothesis, claim, citation*).

We experimented with sentence classification approaches on the Kaggle dataset mentioned above, but found them unsuitable as they do not reflect the gold standard units well. As shown in Figure 2, one sentence can contain multiple argumentative elements, while one argumentative element can span sentences like the *lead*, *counterclaim* and *evidence* annotations in Figure 1. Our sentence classification experiments using a support vector machine reached an F1-Score of only 0.2. We thus did not further proceed with sentence classification on this dataset.



However, driverless cars should be looked at as useful and a positive alternative to everyday cars as they are aware and self-sufficient for their owner's benefit.

Figure 2: An example sentence with multiple argumentative elements from essay 03EA9F90F814 in the Kaggle dataset.

Based on a modification of the Toulmin argument model (Toulmin, 1958), Stab and Gurevych (2014b) proposed a model of argument components in scientific articles and persuasive essays at the **clause**-level using four label types - *major claim, claim, premise* and *non-argumentative*. Their annotation guidelines yielded substantial agreement in an annotation study on 90 persuasive essays in English (Stab and Gurevych, 2014a). Following this schema, the International Corpus of Learner English (Granger et al., 2009) was annotated by Persing and Ng (2015). They trained classification models to identify argument components and used them as features to predict argumentative scores in essays (Persing and Ng, 2016).

In recent research, the granularity of argumentative components was further increased to the **token** level. In this case, the identification of argumentative elements corresponds to assigning an argument label to each word. In this paradigm, sequence labeling techniques like Conditional Random Fields or pretrained BERT models started contributing to argument mining (Trautmann et al., 2020). The Kaggle competition "Feedback Prize - Evaluating Student Writing" can also be seen as a token labeling task as suggested by the organizers[2].

Most studies on argument mining mentioned above do not take the topic of the essay into consideration, assuming that arguments can be classified independently of a topic. However, as shown in studies like Daxenberger et al. (2017), argument mining models did not generalize well on cross-domain data. Subsequently, the importance of topic information has drawn more and more attention in the general argument mining task recently: Stab et al. (2018) found that a topic-general model could achieve comparable performance to a topic-specific model by adding limited amounts of topic-specific data. Fromm et al. (2019) proved that topic information connected with large pretrained language models like BERT provides a significant performance boost in argument mining.

However, the effect of topic information has not been fully examined in educational argument mining. The data released by the Kaggle competition gives us a chance to investigate this research gap, because it not only contains large amounts of student essays with gold standard annotation of different argumentative elements, but also covers essays from a variety of different writing prompts which, while not being annotated in the dataset, can be automatically inferred.

3 Data

As mentioned before, we use the dataset provided as part of the Kaggle competition "Feedback Prize - Evaluating Student Writing". The dataset consists of 15,594 argumentative students essays written by U.S. students from grades 6 to 12. Essays contain annotations for the following argumentative labels:[3]

- Lead: an introduction that begins with a statistic, a quotation, a description, or some other device to grab the reader's attention and point toward the thesis.
- Position: an opinion or conclusion on the main question.
- Claim: a claim that supports the position.
- Counterclaim: a claim that refutes another claim or gives an opposing reason to the position.
- Rebuttal: a claim that refutes a counterclaim.
- Evidence: ideas or examples that support claims, counterclaims, or rebuttals.
- Concluding Statement: a concluding statement that restates the claims.

Argumentative units have been annotated with an overall inter-rater reliability of .73. The lowest reliability was reported for counterclaims and rebuttals (which were often labeled as claims). The highest reliability was found for concluding statements. All disagreements were adjudicated by an expert rater.[4]

Figure 3 shows the frequency and average number of tokens per span for each label in the dataset. We notice that the argumentative components are very unevenly distributed. *Claim* and *evidence* occur substantially more frequently than the other labels, with *counterclaims* and *rebuttals* being particularly rare.

In terms of the length of the underlying span for a label, instances of the types *evidence*, *concluding statement* and *lead* correspond to the longest spans. The average length of all essays is 429 words, while the average length of *evidence* is 77 words, which means that, given the frequency of the label, *evidence* is the majority class on the token level. In contrast, *position* and *claim* have the shortest average length.

3.1 Clustering the Data into Underlying Prompts

The dataset is not annotated with prompt information. To obtain the individual prompts, we first use a topic modeling approach (Angelov, 2020), which resulted in a total of 11 clusters of essays. Manual inspection of a random sample of 25 essays per cluster finds two clusters to be a mixture of either 2 or 4 different prompts. We used a k-means clustering approach on tf-idf vectors per essay to

[2] https://www.kaggle.com/competitions/feedback-prize-2021/discussion/296010

[3] Explanations for the labels are directly taken from the competition webpage: https://www.kaggle.com/competitions/feedback-prize-2021/data

[4] https://www.kaggle.com/c/feedback-prize-2021/discussion/297688

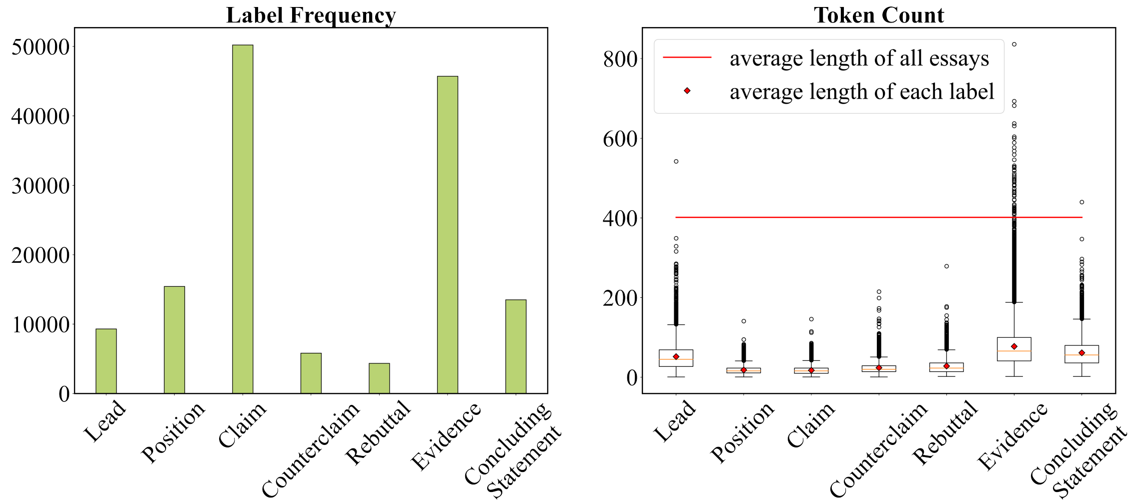


Figure 3: Frequency and token count of labels in the Kaggle Feedback Prize dataset.

further split those clusters into 2 and 4 sub-clusters respectively. The resulting 15 clusters each contain between 689 and 1826 individual essays.

To check the quality of the clusters, we manually annotated 100 instances per cluster and found cluster purity (Manning et al., 2010) to be between 0.78 and 1. Table 1 shows the detected topics and cluster evaluation numbers.

We consider this clustering to be good enough to be used for a topic-based modeling approach without time-intensive manual adjudication of clusters. However, it should be noted that especially for the *"Extracurricular Activities"* cluster with an outlier purity of .78 only, artifacts introduced by impure clustering might occur.

prompt	**#essays**	**purity**
Exploring Venus	930	1.00
Face on Mars	817	1.00
Electoral College	1826	1.00
Phones and Driving	705	.90
Driverless Car	1390	.99
Getting Advice	1414	.99
Phones in School	841	.96
Seagoing Cowboys	689	.97
Summer Projects	860	.98
Facial Action Coding	1055	.99
Community Center	712	1.00
Limiting Car Usage	991	.96
Extracurricular Activities	1146	.78
Online Classes	1457	1.00
No Sports at Grade C	761	1.00

Table 1: Topics detected in the dataset, number of essays per topic and purity of the detected cluster.

4 Experimental Study 1 - The Influence of Prompt Information

In this study, we train argument mining models with different combinations of prompt-specific and cross-prompt data and compare their performance on the same test datasets, in order to investigate our first research question: in how far is argument mining prompt-dependent? Furthermore, we analyze the performance difference among argument labels.

4.1 Experimental Setup

As our base model, we adopt a neural architecture developed for the structurally similar sequence labelling task of Named Entity Recognition (Grishman and Sundheim, 1996). As almost one third of all essays contains more than 512 tokens, we exchange the pretrained BERT token classification model (Devlin et al., 2018) for a pretrained Longformer model (Beltagy et al., 2020) where the attention mechanism scales linearly instead of quadratically with input length. The experiment pipeline is shown in Figure 4. We pre-process the annotated training data into tokens with Inside-Outside-Beginning (IOB) tags and use them as the input to the pretrained Longformer model for token classification (longformer-large-4096). After 10 epochs of training with a maximal length of 1536 tokens, the IOB-Tags of tokens are transformed into predictions for different argumentative elements in the post-processing.

We compare several configurations for the training data: In the **all prompts** condition, we train on the complete dataset with all 15 prompts. In the

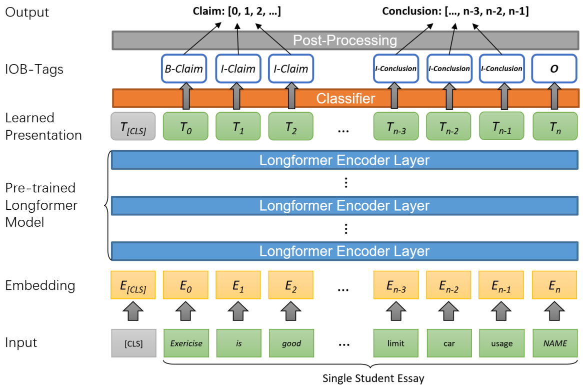


Figure 4: Training pipeline of baseline model.

same prompt condition, we only train on essays from the same prompt as the test data. To create a more controlled setting not influenced by the fraction of essays from the same prompt, we exclude them explicitly in the **other prompts** condition, using 12 of the 14 other prompts for training and 2 for validation. Some prompts in the dataset are closer to each other, as can be seen by the fact that they were confused by our topic clustering approach. To see whether using a **related prompt** is beneficial, we evaluate each of the three prompts *"Driverless Car"*, *"Phones and Driving"* and *"Limiting Car Usage"* under a model trained for either one of the other two prompts.

We split each prompt into 80% training data, 10% evaluation data and 10% test data. To make results across settings comparable, we make sure that we always test on the test data portion only (even for the setting **other prompts** where the whole dataset in a prompt would be available for testing). In order to get comparable models trained on similar amounts of data, we produce another version of the **all prompts** and **other prompts** conditions, where we sample down to the same average amount of training data as used in the **same prompt** condition, called **all prompts – small** and **other prompts – small**.

Following the evaluation scheme proposed by Kaggle, we evaluate based on the overlap between predicted spans and gold standard spans. A prediction is considered a true positive (TP) if the overlap between the prediction and the gold standard is greater than 50% in both directions. Any unmatched ground truths are false negatives (FN), and any unmatched predictions are false positives (FP). The final score is arrived at by calculating TP/FP/FN for each class, then taking the macro F1 score across all classes. Predictions of non-argument text are excluded from the evaluation.

4.2 Experiment 1a - Comparison between Different Training Setups

In our first experiment, we compare the overall performances of the different training setups averaged across all prompts. We cannot use the **related prompt** condition here, as we cannot use all prompts in this condition (simply because not every prompt has a similar other prompt).

Training Data	Avg. Amount Training Data	F1
same prompt	833	.53
other prompts – small	833	.49
all prompts – small	833	.52
other prompts	9983	.52
all prompts	12481	.55

Table 2: Results for Experiment 1a, F1 score averaged over all prompts.

According to the results shown in Table 2, using data from the **same prompt** condition for training brings benefits compared to a setup with the same size of training data drawn from other prompts (**other prompts – small**). **Other prompts** and **all prompts**, in comparison, show the performance on more than 10 times the amount of training data. We observe that using more cross-prompt data (i.e. **other prompts**) provides no advantage compared to fewer data from within the same prompt. However, if some amount of within-prompt data is available, as in **all prompts**, the model benefits from more data. Note that **all prompts** contains all training items from the **same prompt** condition plus material from other prompts. This implies that a prompt-specific model can be slightly improved by adding extra generic data.

4.3 Experiment 1b - Training on Related Prompts

We have seen in Experiment 1a that, given a fixed amount of training data, within-prompt training data from the same prompt is beneficial. However, this can be impractical in a real-life setting, as it might be expensive to obtain new training material for every new essay prompt. Therefore, we investigate in the following experiment whether training on a topic-wise related prompt already helps.

We select three prompts centered around cars and driving: *"Driverless Cars"*, *"Phones and Driving"* and *"Limited Car Usage"*. The fact that these three prompts were often confused during topic clustering shows their relatedness on the lexical level.

Results in Table 3 show that models trained on topic-related data do not have quite the same performance as those trained on data from the same topic or trained on all prompts. The **other prompts – small**, **all prompts – small** and **same prompt** models are the same as in Experiment 1a (but of course only averages over 3 prompts are reported).

Training Data	Avg. Amount Training Data	F1
same prompt	823	.48
other prompts – small	833	.46
all prompts – small	833	.49
related prompt	823	.46

Table 3: Results for Experiment 1b, F1 score averaged over prompts *Driverless Car*, *Phones and Driving* and *Limiting Car Usage*.

4.4 Experiment 2 - Performance Analysis for Individual Argument Labels

As we have seen in Section 3, the dataset is very skewed in terms of the distribution of individual labels. Therefore, we expect the performance of labels with a low frequency a) to be worse than that of more frequent labels and b) to benefit more from larger amounts of training data than the frequent labels.

Results shown in Figure 5 only partially confirm these expectations. We see that performance varies a lot for individual label types, but does not directly reflect the label distribution. While the most infrequent *rebuttal* label also shows the worst classification performance, the labels with the best performance are *lead* and *concluding statement*. Contrary to what we expected, the much more frequent *claim* and *evidence* can be found less precisely, with especially the label *claim* exhibiting the second-lowest performance of all labels.

We speculate that several factors contribute to this behavior. The two argumentation labels with the highest performance are those who potentially benefit most from positional information that a classifier might learn. In the gold standard, 49% of all texts indeed start with a *lead* annotation. If a lead is present in an essay, in 82% of all instances it occurs right in the beginning. Similarly, 70% of all essays end with a *concluding statement* and among all concluding statements, 81% are right at the end of a text. Claims, although very frequent, do not appear at a specific position in the text and are often not clearly marked by discourse markers. We checked the occurrence of a list of about 200 common discourse connectives and discourse markers such as *because*, *although* or *additionally* (Sileo et al., 2019) and found that *counterclaims* and *rebuttals* were most strongly marked by such words - a possible reason why their performance, although these labels are infrequent, is not far below that of claims.

We checked common confusions between labels in our classification results. Table 4 shows that the majority of all confusions occurs between a label and no assigned span, indicating that the assignment of correct argumentation unit boundaries is a problem, which leads to numerous spans with no counterpart with a sufficient overlap. When comparing the number of unmatched gold standard labels (3521) with that of unmatched predicted labels (5781), we see our algorithm tends to assign a label rather than not assign anything. Among the actual confusions between two labels, we observe some confusions also reported for humans, such as counterclaims often being mislabelled as claims.

5 Experimental Study 2 - What do we actually Learn?

Aiming to answer our second research question of whether the algorithms are more susceptible to prompt-specific or general information, we now transform the original data into **topic-only** and **structure-only** versions.

5.1 Experimental Setup

Experimental Study 1 indicated that the identification of argumentative elements benefits from prompt-specific information. However, it remains unclear whether we actually learn to detect topic words constituting, e.g., a typical claim for a certain Topic X or structural elements of a claim in Topic X, which could also be found in other topics. To disentangle the two effects from each other, we perform an additional set of analyses, as detailed in the following.

We filter the vocabulary according to how often it appears within a specific prompt and in the overall dataset. Similar to a tf-idf approach (Ramos et al., 2003), we consider vocabulary prompt-specific if it appears often within the essays of one prompt, but infrequently within the essays of other

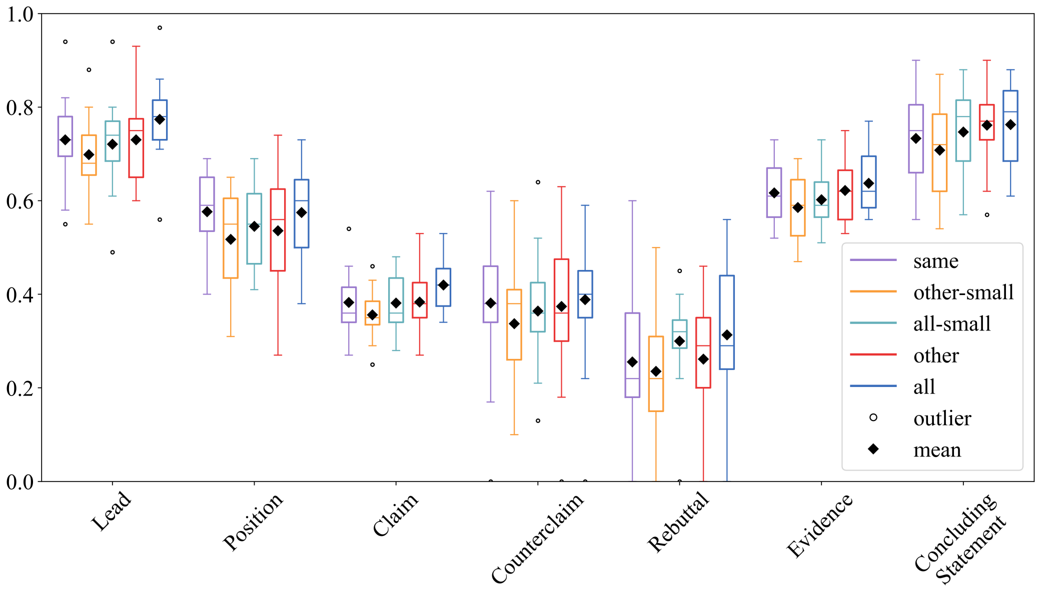


Figure 5: Results for Experiment 2, F1 scores of each label in different settings

	Lead	**Position**	**Claim**	**Counter-claim**	**Rebuttal**	**Evidence**	**Conclu. Statem.**	**None**
Lead	715	35	2	0	0	17	0	135
Position	38	765	12	6	2	11	2	578
Claim	8	75	1659	4	17	153	4	2912
Conterclaim	2	6	33	264	2	19	1	268
Rebuttal	0	4	19	4	149	23	2	234
Evidence	22	17	242	39	66	2934	29	1449
Conclu. Statem.	0	39	21	4	4	42	1098	205
None	147	245	1145	157	144	1388	302	N.A.

Table 4: Confusion matrix between gold standard (columns) and results in the *same prompt* setting (rows)

prompts. For example, the word *Mars* appears 7851 times in the *"Face on Mars"* prompt, but only 448 times in all other prompts. We rank word types in each prompt by their tf-idf value and consider the top 1000 types as the **topic words** of each prompt.

We then produce 4 versions of the data. In the **structure-only** versions, topic words in each prompt are replaced by the mask word *"dummy"* (**structure-only-dummy**) or their part-of-speech (POS) tags (**structure-only-pos**). The usage of POS tags is intended to keep the syntactic structure intact. In the complementary **topic-only** versions, every occurrence of any non-topical words as well as every function word is replaced by the dummy word (**topic-only-dummy**) or its POS tag (**topic-only-pos**). Table 5 shows an example for the resulting sentences.

We now perform scoring experiments comparable to those from Experimental Study 1 on the modified data. Similar to a feature ablation test, we want to examine how masking some part of the information present in an essay affects the classification outcome.

5.2 Experiment 3a - Modified Test Data

In this experiment, we use the method described above to modify only the test data (the same 10% test data used in Experiment 1). We compare the prediction of models from Experimental Study 1 trained in the settings **same prompt** and **other prompts – small** on the modified data in order to test what kind of information the models have learned. We hypothesize that the **same prompt** model learns both prompt-related and generic structural information, while **other prompts – small** - in the absence of prompt-specific information - learns only general structure as predictor for argumentative elements.

The results shown as orange bars in Figure 6 reveal that, unsurprisingly, the general performance of models is much lower than the performance on the original test data. Nevertheless, we see

Version	Sentence
Original	The Face on Mars is a natural landform
Structure-Only-Dummy	The dummy on dummy is a dummy dummy
Structure-Only-Pos	The [NNP] on [NNP] is a [JJ] [NN]
Topic-Only-Dummy	dummy Face dummy Mars dummy dummy natural landform
Topic-Only-Pos	[DT] Face [IN] Mars [VBZ] [DT] natural landform

Table 5: Four versions of one sentence generated according to our four individual conditions.

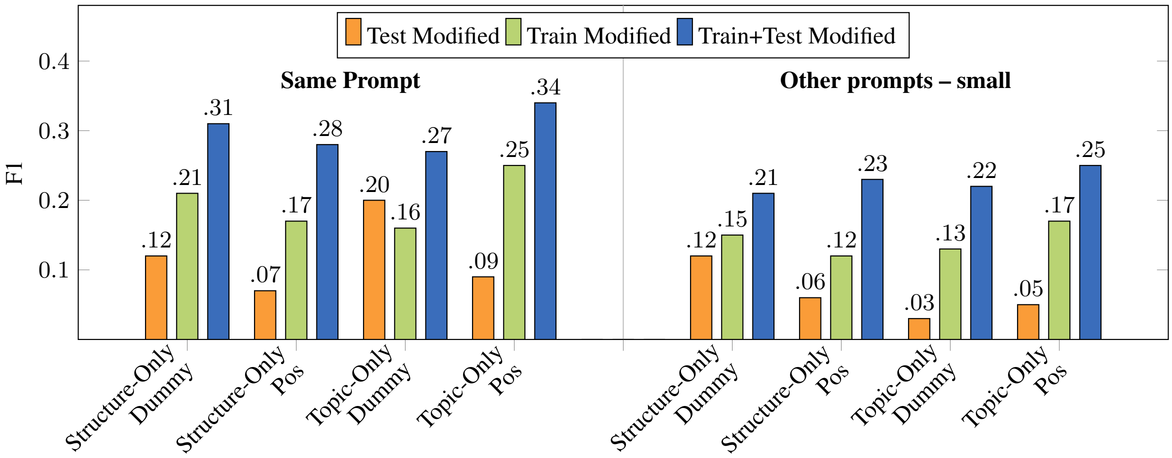


Figure 6: Performance of using the model trained on original data to evaluate modified testing data (test modified), training a model on modified data and testing it on the original testing data (train modified), or both training and testing on modified data.

that the models trained on the **same prompt** perform better on topic-only data than structure-only data. In the **other prompts – small** setting, in contrast, structure-only training data works better than topic-only data, indicating that those models indeed mainly learn structural information.

5.3 Experiment 3b - Modified Training Data

Similar to the **same prompt** and **other prompts – small** settings in Experimental Study 1, we train two models for each prompt on each of the four modified versions of the data. By applying these models to the original test data, we get the results shown as green bars in Figure 6.

Among all models, we expect models trained on topic-only data from **other prompts – small** to have the worst performance, since the predictors learned in these models are theoretically only content words related to other topics. However, the models trained on topic-only data have comparable performance to other models in the **other prompts – small** setting, a fact that needs further investigation and that might be due to either impure clusters or content word filtering (such that the training data still contains some usable lexical information), or to the fact that positional information is a strong predictor present in all our modified data variants.

In the **same prompt** setting, models trained on topic-only-pos data also have the best performance. But once the POS-tags are changed into *"dummy"* (i.e. topic-only-dummy), the models cannot beat those trained on structure-only data.

5.4 Experiment 3c - Modified Training and Test Data

Finally, we use the models trained in Experiment 3b on modified data and test on modified test data as well. Results are shown in Figure 6 as blue bars. Unsurprisingly, these models with train and test data modified in the same way yield better performance compared to those where only the train or the test data was modified and, similar to the results above, models trained on data from **same-prompt** perform better than those trained on data from **other prompts – small** in general. They still perform far below the level of the original experiments, indicating that in both conditions, models benefit from both structural and topical information. However, the loss is larger in the **other** conditions than for **same**.

Similar to the results in Experiment 3b, the models trained on topic-only-pos in the **same-prompt** setting have the best performance, because not only topic related information is kept in the training data, but also limited structural information is included by the POS-tags.

6 Conclusion

This work set out to investigate the importance of topic information in educational argument mining tasks. For this purpose, we first clustered a recently published dataset of student essays into underlying prompts. Secondly, we presented a study on the effect of prompt-specific and cross-prompt training material in the identification of argumentative elements. Results showed that within-prompt training data is beneficial when a fixed limited amount of training data is used. This advantage can be overcome by larger amounts of additional cross-prompt data. In the analysis of argumentative elements, we found that *lead* and *conclusion* can be best identified in all settings, presumably because of their relatively fixed position. Lastly, we separated topical from structural information in the essays. From experiments with this modified data, we found that argument mining benefits both from topic words and structure words, i.e. the information is not redundant, but that, unsurprisingly, topical information has a tendency to be more important in within-prompt classification while structure is more relevant across prompts.

These findings provide the following insights for future research: first, learning curve studies could investigate an optimal trade-off between topic-specific and generic training data. Second, the argumentative elements identified in student essays could be meaningful for the generation of formative feedback directed towards students, such as highlighting different argumentative elements. Another research direction is the evaluation of argument quality through analyzing discourse relations between these argument components in order to generate feedback towards coherence and cohesion aspects of student essays.

Acknowledgements

This work was conducted in the framework of the Research Cluster D^2L^2 “Digitalization, Diversity and Lifelong Learning – Consequences for Higher Education” of the FernUniversität in Hagen, Germany [5]. We thank our student assistants Finn Brodmann and Joey Pehlke (University of Duisburg-Essen) who contributed to data preparation and analysis.

References

Dimo Angelov. 2020. Top2Vec: Distributed Representations of Topics. *arXiv e-prints* pages arXiv–2008.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The Long-document Transformer. *arXiv preprint arXiv:2004.05150* .

Jill Burstein, Daniel Marcu, and Kevin Knight. 2003. Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. *IEEE Intelligent Systems* 18(1):32–39.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the Essence of a Claim? Cross-Domain Claim Identification. *arXiv preprint arXiv:1704.07203* .

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* .

Michael Fromm, Evgeniy Faerman, and Thomas Seidl. 2019. TACAM: Topic And Context Aware Argument Mining. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, pages 99–106.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, Magali Paquot, et al. 2009. *International Corpus of Learner English*, volume 2. Presses universitaires de Louvain Louvain-la-Neuve.

Ralph Grishman and Beth M Sundheim. 1996. Message Understanding Conference-6: A Brief History. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

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

Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. 2010. Introduction to Information Retrieval. *Natural Language Engineering* 16(1):100–103.

Raquel Mochales and Aagje Ieven. 2009. Creating an Argumentation Corpus: do Theories Apply to Real Arguments? A Case Study on the Legal Argumentation of the ECHR. In *Proceedings of the 12th international conference on artificial intelligence and law*. pages 21–30.

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.

[5] https://e.feu.de/english-d2l2

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 (Volume 1: Long Papers)*. pages 543–552.

Isaac Persing and Vincent Ng. 2016. End-to-End Argumentation Mining in Student Essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 1384–1394.

Juan Ramos et al. 2003. Using TF-IDF to Determine Word Relevance in Document Queries. In *Proceedings of the first instructional conference on machine learning*. Citeseer, volume 242, pages 29–48.

Oliver Scheuer, Frank Loll, Niels Pinkwart, and Bruce M McLaren. 2010. Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-supported collaborative learning* 5(1):43–102.

Damien Sileo, Tim Van De Cruys, Camille Pradel, and Philippe Muller. 2019. Mining Discourse Markers for Unsupervised Sentence Representation Learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, pages 3477–3486.

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.

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.

Christian Stab and Iryna Gurevych. 2017. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics* 43(3):619–659.

Christian Stab, Tristan Miller, and Iryna Gurevych. 2018. Cross-Topic Argument Mining from Heterogeneous Sources using Attention-based Neural Networks. *arXiv preprint arXiv:1802.05758* .

Stephen E Toulmin. 1958. The Uses of Argument .

Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2020. Fine-Grained Argument Unit Recognition and Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*. volume 34, pages 9048–9056.

Marilyn Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A Corpus for Research on Deliberation and Debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. pages 812–817.

Ursula Wingate. 2012. 'Argument!' helping Students understand What Essay Writing is about. *Journal of English for academic purposes* 11(2):145–154.