# Recognizing the Absence of Opposing Arguments in Persuasive Essays

**Christian Stab**[†] and **Iryna Gurevych**[†‡]
[†]Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
[‡]Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research
www.ukp.tu-darmstadt.de

## Abstract

In this paper, we introduce an approach for recognizing the absence of opposing arguments in persuasive essays. We model this task as a binary document classification and show that adversative transitions in combination with unigrams and syntactic production rules significantly outperform a challenging heuristic baseline. Our approach yields an accuracy of $75.6\%$ and $84\%$ of human performance in a persuasive essay corpus with various topics.

## 1 Introduction

Developing well-reasoned arguments is an important ability and constitutes an important part of education programs (Davies, 2009). A frequent mistake when writing argumentative texts is to consider only arguments supporting the own standpoint and to ignore opposing arguments (Wolfe and Britt, 2009). This tendency to ignore opposing arguments is known as *myside bias* or *confirmation bias* (Stanovich et al., 2013). It has been shown that guiding students to include opposing arguments in their writings significantly improves the argumentation quality, the precision of claims and the elaboration of reasons (Wolfe and Britt, 2009). Therefore, it is likely that a system which automatically recognizes the absence of opposing arguments effectively guides students to improve their argumentation. For the same reason, the writing standards of the *common core standard*[1] require that students are able to clarify the relation between their own standpoint and opposing arguments on a controversial topic.

Existing *structural approaches* on argument analysis like the argumentation structure parser presented by Stab and Gurevych (2016) or the approach introduced by Peldszus and Stede (2015a) recognize the internal microstructure of arguments. Although these approaches can be exploited for identifying opposing arguments, they require several consecutive analysis steps like separating argumentative from non-argumentative text units (Moens et al., 2007), recognizing the boundaries of argument components (Goudas et al., 2014) and classifying individual arguments as support or oppose (Somasundaran and Wiebe, 2009). Certainly, an advantage of structural approaches is that they recognize the position of opposing arguments in text. However, knowing the position of opposing arguments is only relevant for positive feedback to the author and irrelevant for negative feedback, i.e. pointing out that opposing arguments are missing. Therefore, it is reasonable to model the recognition of missing opposing arguments as a document classification task.

The contributions of this paper are the following: first, we introduce a corpus for detecting the absence of opposing arguments that we derive from argument structure annotated essays. Second, we propose a novel model and a new feature set for detecting the absence of opposing arguments in persuasive essays. We show that our model significantly outperforms a strong heuristic baseline and an existing structural approach. Third, we show that our model achieves $84\%$ of human performance.

## 2 Related Work

Existing approaches in computational argumentation focus primarily on the identification of arguments, their components (e.g. claims and premises) (Rinott et al., 2015; Levy et al., 2014) and structures (Mochales-Palau and Moens, 2011; Stab and Gurevych, 2014b). Among these, there

[1] www.corestandards.org



*Proceedings of the 3rd Workshop on Argument Mining*, pages 113–118,
Berlin, Germany, August 7-12, 2016. 

are few approaches which distinguish between supporting and opposing arguments.

Peldszus and Stede (2015b) use lexical, contextual and syntactic features to classify argument components as support or oppose. They experiment with pro/contra columns of a German newspaper and German microtexts. Similarly, their minimum spanning tree (MST) approach identifies the structure of arguments and recognizes if an argument component belongs to the proponent or opponent (Peldszus and Stede, 2015a). However, both approaches presuppose that the components of an argument are already known. Thus, they omit important analysis steps and cannot be applied directly for recognizing the absence of opposing arguments. Stab and Gurevych (2016) present an argumentation structure parser that includes all required steps for identifying argument structures and supporting and opposing arguments. First, they separate argumentative from non-argumentative text units using conditional random fields (CRF). Second, they jointly model the argument component types and argumentative relations using integer linear programming (ILP) and finally they distinguish between supporting and opposing arguments. We employ this parser as a structural approach and compare it to our document classification approach for recognizing the absence of opposing arguments in persuasive essays.

Another related area is *stance recognition* that aims at identifying the author's stance on a controversy by labeling a document as either "for" or "against" (Somasundaran and Wiebe, 2009; Hasan and Ng, 2014). Consequently, stance recognition systems are designed to identify the predominant stance of a text instead of recognizing the presence of less conspicuous opposing arguments.

Other approaches on argumentation in essays focus on thesis clarity (Persing and Ng, 2013), argumentation schemes (Song et al., 2014) or argumentation strength (Persing and Ng, 2015). We are not aware of any approach that focuses on recognizing the absence of opposing arguments.

## 3 Data

For our experiments, we employ an argument structure annotated essay corpus (Stab and Gurevych, 2014a; Stab and Gurevych, 2016). To the best of our knowledge, this corpus is the only available resource that exhibits an appropriate size and class distribution for detecting the absence of opposing arguments at the document-level. Each essay in this corpus is annotated with argumentation structures that allow to derive document-level annotations. The argumentation structures include arguments supporting or opposing the author's stance. Accordingly, we consider an essay as *negative* if it solely includes supporting arguments and as *positive* if it includes at least one opposing argument. Note that the manual identification of opposing arguments is a subtask of the argumentation structure identification. Both require that the annotators identify the author's stance, the individual arguments and if an argument supports or opposes the author's stance. Thus, deriving document-level annotations from argumentation structures is a valid approach since the decisions of the annotators in both tasks are equivalent.

### 3.1 Inter-Annotator Agreement

To verify that the derived document-level annotations are reliable, we compare the annotations derived from the argumentation structure annotations of three independent annotators. In particular, we determine the inter-annotator agreement on a subset of 80 essays. The comparison shows an observed agreement of $90\%$. We obtain substantial chance-corrected agreement scores of Fleiss' $\kappa = .786$ (Fleiss, 1971) and Krippendorff's $\alpha = .787$ (Krippendorff, 2004). Thus, we conclude that the derived annotations are reliable since they are only slightly below the "*good reliability threshold*" proposed by Krippendorff (2004).

### 3.2 Statistics

Table 1 shows an overview of the corpus. It includes 402 essays. On average each essay includes 18 sentences and 366 tokens.

| | |
|---|---|
| Tokens | 147,271 |
| Sentences | 7,116 |
| Documents | 402 |
| Negative | 251 (62.4%) |
| Positive | 151 (37.6%) |

Table 1: Size and class distribution of the corpus.

The class distribution is skewed towards negative essays. The corpus includes 251 ($62.4\%$) essays that do not include opposing arguments and 151 ($37.6\%$) positive essays. For encouraging future research, the corpus is freely available.[2]

[2]https://www.ukp.tu-darmstadt.de/data

## 4 Approach

We consider the recognition of opposing arguments as a binary document classification. Due to the size of the corpus and to prevent errors in model assessment stemming from a particular data splitting (Krstajic et al., 2014), we employ a stratified and repeated 5-fold cross-validation setup. We report the average evaluation scores and the standard deviation over 100 folds resulting from 20 iterations. For model selection, we randomly sampled 10% of the training set of each run as a development set. We report accuracy, macro precision, macro recall and macro F1 scores as described by Sokolova and Lapalme (2009, p. 430).[3] We employ Wilcoxon signed-rank test on macro F1 scores for significance testing (significance level = .005).

We preprocess the essays using several models from the DKPro framework (Eckart de Castilho and Gurevych, 2014). For tokenization, sentence and paragraph splitting, we employ the language tool segmenter[4] and check for line breaks. We lemmatize each token using the mate tools lemmatizer (Bohnet et al., 2013) and apply the Stanford parser (Klein and Manning, 2003) for constituency and dependency parsing. Finally, we use a PDTB parser (Lin et al., 2014) and sentiment analyzer (Socher et al., 2013) for identifying discourse relations and sentence-level sentiment scores. As a learner, we choose a support vector machine (SVM) (Cortes and Vapnik, 1995) with polynomial kernel implemented in Weka (Hall et al., 2009). For extracting features, we use the DKPro TC framework (Daxenberger et al., 2014).

### 4.1 Features

We experiment with the following features:

**Unigrams (uni):** In order to capture the lexical characteristics of an essay, we extract binary and case sensitive unigrams.

**Dependency triples (dep)**: The binary dependency features include triples consisting of the lemmatized governor, the lemmatized dependent and the dependency type.

**Production rules (pr)**: We employ binary production rules extracted from the constituent parse trees (Lin et al., 2009) that occur at least five times.

**Adversative transitions (adv)**: We assume that opposing arguments are frequently signaled by lexical indicators. We use 47 adversative transitional phrases that are compiled as a learning resource[5] and grouped in the following categories: concession (18), conflict (12), dismissal (9), emphasis (5) and replacement (3). For each of the five categories, we add two binary features set to true if a phrase of the category is present in the surrounding paragraphs (introduction or conclusion) or in a body paragraph.[6] Note that we consider lowercase and uppercase versions of these features which results in a total of 20 binary features.

**Sentiment Features (sent)**: We average the five sentiment scores of all essay sentences for determining the global sentiment of an essay. In addition, we count the number of negative sentences and define a binary feature indicating the presence of a negative sentence.

**Discourse relations (dis)**: The binary discourse features include the type of the discourse relation and indicate if the relation is implicit or explicit. For instance, "*Contrast_imp*" indicates an implicit contrast relation. Note that we only consider the discourse relations of body paragraphs since the introduction frequently includes a description of the controversy which is not relevant to the author's argumentation and whose discourse relations could be misleading for the learner.

### 4.2 Baselines

For model assessment, we use the following two baselines: First, we employ a *majority baseline* that classifies each essay as negative (not including opposing arguments). Second, we employ a rule-based *heuristic baseline* that classifies an essay as positive if it includes the case-sensitive term "*Admittedly*" or the phrase "*argue that*" which often indicate the presence of opposing arguments.[7]

### 4.3 Results

In order to select a model and to analyze our features, we conduct feature ablation tests (lower part of Table 2) and evaluate our system with individual features. The adversative transitions and unigrams are the most informative features. Both show the best individual performance and a sig-

[3] Since the macro F1 score assigns equal weight to classes, it is well-suited for evaluating experiments with skewed data.

[4] www.languagetool.org

[5] www.msu.edu/~jdowell/135/transw.html

[6] We identify paragraphs by checking for line breaks and consider the first paragraph as introduction, the last as conclusion and all remaining ones as body paragraphs.

[7] We recognized these indicators by ranking n-grams using information gain.

| | *Accuracy* | *Macro F1* | *Precision* | *Recall* | *F1 Negative* | *F1 Positive* |
|---|---|---|---|---|---|---|
| *Model assessment on test data* | | | | | | |
| Human Upper Bound* | .900±.010 | .894±.011 | .895±.011 | .014±.892 | .865±.016 | .921±.008 |
| Baseline Majority | .624±.001 | .384±.000 | .312±.001 | .500±.000 | .769±.001 | 0 |
| Baseline Heuristic | .711±.039 | .679±.050 | .715±.059 | .646±.045 | .797±.027 | .497±.083 |
| SVM uni+pr+adv † | **.756**±.044 | **.734**±.048 | **.747**±.049 | **.721**±.050 | **.814**±.034 | **.639**±.075 |
| *Model selection and feature ablation on development data* | | | | | | |
| SVM all w/o uni ‡ | .733±.060 | .708±.087 | .768±.110 | .660±.073 | .817±.038 | .496±.151 |
| SVM all w/o dep | .765±.077 | .745±.087 | .762±.092 | .731±.086 | .822±.059 | .649±.125 |
| SVM all w/o pr | .760±.062 | .738±.082 | **.781**±.097 | .701±.074 | **.830**±.042 | .583±.138 |
| SVM all w/o adv ‡ | .736±.066 | .709±.090 | .756±.108 | .670±.079 | .816±.044 | .524±.151 |
| SVM all w/o sent | .756±.064 | .733±.085 | .778±.100 | .696±.076 | .828±.043 | .572±.146 |
| SVM all w/o dis | .757±.061 | .734±.082 | .780±.097 | .696±.075 | .829±.041 | .571±.143 |
| SVM uni+pr+adv | **.770**±.071 | **.750**±.081 | .767±.086 | **.735**±.080 | .825±.055 | **.656**±.118 |
| SVM all features | .755±.064 | .732±.086 | .776±.102 | .695±.077 | .827±.044 | .569±.149 |

Table 2: Results of the best performing model on the test data and selected results of the model selection experiments on the development data († significant improvement over *Baseline Heuristic*; ‡ significant difference compared to *SVM all features*; *determined on a subset of 80 essays).

nificant decrease if removed from the entire feature set. Thus, we conclude that lexical indicators are the most predictive features in our feature set. The sentiment and discourse features do not perform well. Individually they do not achieve better results than the majority baseline and the accuracy increases slightly when removing them from the entire feature set. By experimenting with various feature combinations, we found that combining unigrams, production rules and adversative transitions yields the best results (*SVM uni+pr+adv*).

For model assessment, we evaluate the best performing model on our test data and compare it to the baselines (upper part of Table 2). The heuristic baseline considerably outperforms the majority baseline and achieves an accuracy of 71.1%. Our best system significantly outperforms this challenging baseline with respect to all evaluation measures. It achieves an accuracy of 75.6% and a macro F1 score of .734. We determine the human upper bound by comparing pairs of annotators and averaging the results of the 80 independently annotated essays (cf. Section 3). Compared to the upper bound, our system achieves 14.4% less accuracy and 84% of human performance.

We compare our system to an argumentation structure parser that recognizes opposing components on a designated 80:20 train-test-split (Stab and Gurevych, 2016). We consider essays with predicted opposing arguments as positive, and negative if the parser does not recognize an opposing argument. This yields a macro F1 score of .648. Our document-level approach considerably outperforms the component-based approach with a macro F1 score of .710. Thus, we can confirm our assumption that modeling the task as document classification outperforms structural approaches.

### 4.4 Error Analysis

To analyze frequent errors of our system, we manually investigate essays that are misclassified in all 100 runs of the repeated cross-validation experiment on the development set. In total, 29 positive essays are consistently misclassified as negative. As reason for these errors, we found that the opposing arguments in these essays lack lexical indicators. In addition, we found 14 negative essays which are always misclassified as positive. Among these essays, we observe that the majority includes opposition indicators (e.g. "*but*") which are used in another sense (e.g. expansion). Therefore, the investigation of both false negatives and false positives shows that most errors are due to misleading lexical signals. Consequently, word-sense disambiguation for identifying senses or the integration of domain and world knowledge in the absence of lexical signals could further improve the results.

## 5 Conclusion

We introduced the novel task of recognizing the absence of opposing arguments in persuasive essays. In contrast to existing structural approaches, we model this task as a document classification which does not presuppose several complex analysis steps. The analysis of several features showed that adversative transitions and unigrams are most indicative for this task. We showed that our best model significantly outperforms a strong heuristic baseline, yields a promising accuracy of 75.6%,

outperforms a structural approach and achieves 84% of human performance. For future work, we plan to integrate the system in writing environments and to investigate its effectiveness for fostering argumentation skills.

## Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806 and by the German Federal Ministry of Education and Research (BMBF) as a part of the Software Campus project AWS under grant No. 01|S12054. We thank Anshul Tak for his valuable contributions.

## References

Bernd Bohnet, Joakim Nivre, Igor Boguslavsky, Richárd Farkas, Filip Ginter, and Jan Hajič. 2013. Joint morphological and syntactic analysis for richly inflected languages. *Transactions of the Association for Computational Linguistics*, 1:415–428.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Peter Davies. 2009. Improving the quality of students' arguments through 'assessment for learning'. *Journal of Social Science Education (JSSE)*, 8(2):94–104.

Johannes Daxenberger, Oliver Ferschke, Iryna Gurevych, and Torsten Zesch. 2014. DKPro TC: A Java-based framework for supervised learning experiments on textual data. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. System Demonstrations*, ACL '14, pages 61–66, Baltimore, MD, USA.

Richard Eckart de Castilho and Iryna Gurevych. 2014. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT) at COLING 2014*, pages 1–11, Dublin, Ireland.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2014. Argument extraction from news, blogs, and social media. In *Artificial Intelligence: Methods and Applications*, volume 8445 of *Lecture Notes in Computer Science*, pages 287–299. Springer International Publishing.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 751–762, Doha, Qatar.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Sapporo, Japan.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to its Methodology*. Sage, 2nd edition.

Damjan Krstajic, Ljubomir J. Buturovic, David E. Leahy, and Simon Thomas. 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics*, 6(10):1–15.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of the 25th International Conference on Computational Linguistics*, COLING '14, pages 1489–1500, Dublin, Ireland.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, EMNLP '09, pages 343–351, Stroudsburg, PA, USA.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Raquel Mochales-Palau 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*, ICAIL '07, pages 225–230, Stanford, CA, USA.

Andreas Peldszus and Manfred Stede. 2015a. Joint prediction in mst-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, pages 938–948, Lisbon, Portugal.

Andreas Peldszus and Manfred Stede. 2015b. Towards detecting counter-considerations in text. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 104–109, Denver, CO.

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)*, ACL '13, pages 260–269, Sofia, Bulgaria.

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)*, ACL '15, pages 543–552, Beijing, China.

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*, EMNLP '15, pages 440–450, Lisbon, Portugal.

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*, EMNLP '13, pages 1631–1642, Seattle, WA, USA.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, ACL '09, pages 226–234, Suntec, Singapore.

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, Baltimore, MA, USA.

Christian Stab and Iryna Gurevych. 2014a. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics*, COLING '14, pages 1501–1510, Dublin, Ireland.

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 '14, pages 46–56, Doha, Qatar.

Christian Stab and Iryna Gurevych. 2016. Parsing argumentation structures in persuasive essays. *arXiv preprint arXiv:1604.07370*.

Keith E. Stanovich, Richard F. West, and Maggie E. Toplak. 2013. Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, 22(4):259–264.

Christopher R. Wolfe and M. Anne Britt. 2009. Argumentation schema and the myside bias in written argumentation. *Written Communication*, 26(2):183–209.