Intrinsic Quality Assessment of Arguments

Henning Wachsmuth
Department of Computer Science
Paderborn University
Paderborn, Germany
henningw@upb.de

Till Werner
Department of Computer Science
Paderborn University
Paderborn, Germany
wtill@mail.upb.de

Abstract

Several quality dimensions of natural language arguments have been investigated. Some are likely to be reflected in linguistic features (e.g., an argument's arrangement), whereas others depend on context (e.g., relevance) or topic knowledge (e.g., acceptability). In this paper, we study the *intrinsic* computational assessment of 15 dimensions, i.e., only learning from an argument's text. In systematic experiments with eight feature types on an existing corpus, we observe moderate but significant learning success for most dimensions. Rhetorical quality seems hardest to assess, and subjectivity features turn out strong, although length bias in the corpus impedes full validity. We also find that human assessors differ more clearly to each other than to our approach.

1 Introduction

Good arguments help to persuade people, to compromise, or to at least understand each other better. What quality dimension is meant by *good* depends on the setting, though (van Eemeren and Grootendorst, 2004; Johnson and Blair, 2006). Several dimensions may be assessed computationally, as we exemplify for an argument in favor of advancing the common good, taken from Wachsmuth et al. (2017a):

> *"While striving to make advancements for the common good you can change the world forever. Allot of people have succeded in doing so. Our founding fathers, Thomas Edison, George Washington, Martin Luther King jr, and many more. These people made huge advances for the common good and they are honored for it."*

The argument is *well-organized* (Persing et al., 2010), its premises are certainly largely *acceptable* (Yang et al., 2019) and *relevant* to the topic (Wachsmuth et al., 2017c). Whether they *suffice* to draw the conclusion (Stab and Gurevych, 2017) is another question, let alone how *convincing* the argument is (Habernal and Gurevych, 2016b). Some dimensions may be reflected in linguistic features of an argument's text. Others depend on context, require topic or background knowledge, or are inherently subjective.

In this paper, we benchmark what quality dimensions of an argument can be assessed intrinsically, i.e., based on the argument's text only. Given a corpus with 304 English arguments on 16 topics scored for 15 dimensions by three experts (Wachsmuth et al., 2017b), we carry out systematic leave-one-topic-out cross-validation experiments. We learn supervised score regression on various text features; from content and distributional semantics, to style, structure, and length, to text quality, evidence, and subjectivity.

For 11 dimensions, we observe moderate but significant gains over a mean baseline. Following intuition, rhetorical quality related to credibility and emotions seem hardest to assess. Capturing subjectivity (e.g., sentiment and pronoun usage) turns out particularly effective. Even better performs the length feature, though, revealing bias in the corpus and matching previous findings on other corpora (Potash et al., 2017). Follow-up experiments indicate that experts strongly differ in assessability, and that some beat our features only slightly. Altogether, an intrinsic assessment of argument quality seems useful but not enough alone.

2 Related Work

Soon after the rise of argument mining, argument quality assessment has come up as a task (Stede and Schneider, 2018), due to its importance for applications such as *Project Debater* (Gleize et al., 2019). It rests on extensive theoretical discussions about what good arguments (Johnson and Blair, 2006) and bad

arguments are (Walton, 2006), and how to argue reasonably (van Eemeren and Grootendorst, 2004).

Several corpora and approaches were proposed for specific argument quality dimensions, first related to essay scoring (Persing and Ng, 2015), some of which modeling arguments explicitly (Wachsmuth et al., 2016). Later approaches targeted arguments from debate portals (Wei et al., 2016), student essays (Stab and Gurevych, 2017), mixed web texts (Wachsmuth et al., 2017c), and news editorials (Yang et al., 2019; El Baff et al., 2020). We use the corpus of Wachsmuth et al. (2017b), as it is the only one annotated for diverse dimensions and is claimed to reflect argument quality comprehensively. In follow-up work, Wachsmuth et al. (2017a) found correlations with the convincingness reasons of Habernal and Gurevych (2016a), and Potthast et al. (2019) as well as Gretz et al. (2020) have evaluated their annotations against the quality annotation scheme of the corpus. However, we do not know any previous assessment approach developed on the corpus, possibly due to its limited size (see Section 3).

We focus on features intrinsic to an argument's text. This complements the study of Potash et al. (2017) who employed external knowledge to assess convincingness. Like them, we find that longer arguments tend to be judged better. Toledo et al. (2019) limit arguments to at most 36 words, avoiding length bias but also preventing deeper reasoning. Quality in their corpus reflects which argument is in doubt preferred. Ultimately, such judgments remain subjective (Lukin et al., 2017). To alleviate this, El Baff et al. (2018) encode the reader's ideology and personality, but such information is often not given in practice.

3 Data

The corpus of Wachsmuth et al. (2017b) is a subset of 320 debate portal posts from the dataset of Habernal and Gurevych (2016a), 20 each for 16 controversial topics. Three human experts (all authors of the paper) scored all posts that they saw as arguments for the following 15 logical, rhetorical, and dialectical quality dimensions on a scale from 1 (low) to 3 (high). In line with the experiments of Wachsmuth et al. (2017a), we use only those 304 texts in Section 5 that were seen as arguments by all three experts.

Logic The main logical dimension is called *cogency (Cog)*. It is defined based on three subdimensions: the *local acceptability (LAc)* of the truth of an argument's premises, the premises' *local relevance (LRe)* for the argument's conclusion, and their *local sufficiency (LSu)* to infer the conclusion.

Rhetoric The main rhetorical dimension is the *effectiveness (Eff)* in persuading readers. Subdimensions are the argument's *clarity (Cla)*, the author's *credibility (Cre)*, the *appropriateness (App)* of the argument's used language, its success in *emotional appeal (Emo)*, and its sequential *arrangement (Arr)* in the text.

Dialectic The main dialectical dimension is *reasonableness (Rea)*, with three subdimensions: the *global acceptability (GAc)* of stating the argument when discussing a given issue, the argument's *global relevance (GRe)* for achieving agreement, and its *global sufficiency (GSu)* in discussing both sides of the issue.

Overall The *overall quality (OvQ)* reflects the subjective weighting of all 14 other quality dimensions.

Both the single expert scores and the mean score are provided for each dimension. The inter-annotator agreement for the different dimensions ranged from 0.26 (emotional appeal) to 0.51 (overall quality) in terms of Krippendorff's α. The majority score of most dimensions is 2, but both score 1 and 3 also occur frequently for many of them. Matching the hierarchical idea of the dimensions, overall quality correlates strongest with cogency, effectiveness, and reasonableness. For details on agreements, score distributions, and correlations, we refer the reader to the original paper (Wachsmuth et al., 2017b).

4 Approach

This paper does *not* aim to propose a novel quality assessment approach, but to evaluate what features of a text help to assess which quality dimension. External knowledge is included only via lexicons and embeddings. As an example, Figure 1 shows selected textual aspects of the argument from Section 1 that may be predictive of certain dimensions. We quantify these and other aspects in the following eight feature types that are employed in linear SVMs for score regression (Chang and Lin, 2011):[1]

[1]We also tested some pretrained configurations of BERT (Devlin et al., 2018), fine-tuning them during training. However, performance was low, possibly due to the small data and the prevention of learning topic information in our experimental setup.

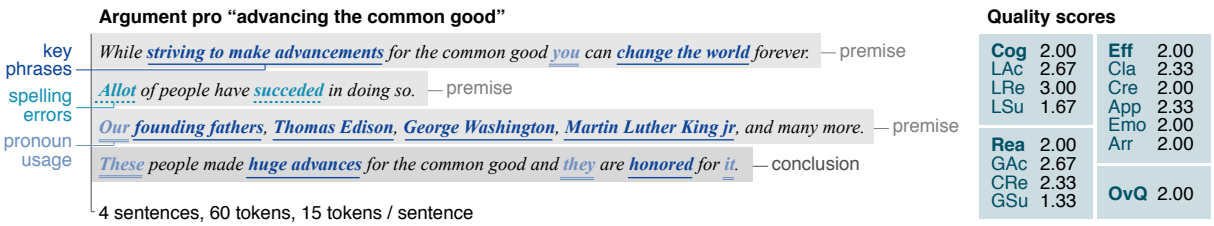

Figure 1: Exemplary analysis of an argument from the used corpus for selected text features that might affect quality: content-related *key phrases*, text quality indicators such as *spelling errors*, subjective *pronoun usage*, length in *sentences/tokens*, and evidence distributions reflected by *premises* and *conclusions*. On the right, all mean quality scores of the argument in the corpus are shown (worst is 1, best is 3).

Content As often done in text classification (Aggarwal and Zhai, 2012), we aim to capture important content-related key phrases simply as part of the distribution of word 1- to 3-grams, taking all those that occur in $\geq 3\%$ of all training texts (such thresholds were set after initial tests).

Embedding We capture an argument's distributional semantics by a sentence vector, using the pretrained *fasttext* model based on Wikipedia (Mikolov et al., 2018). Each vector position becomes one feature.

Style We model style with part-of-speech 1- to 3-grams ($\geq$10% training frequency) and character 1- to 3-grams ($\geq$3%). Both are common stylometric features in authorship attribution (Stamatatos, 2009).

Structure In terms of structure, we look for enumeration indicators, such as “1.” and “2.” or “on one hand” and “on the other hand”. In addition, we check the first token 1-, 2-, and 3-gram in the text.

Length Our length feature type includes normalized counts of characters, syllables, tokens, phrases, sentences, and paragraphs, along with ratios between each pair of these linguistic units.

Text Quality Motivated by classical essay scoring (Ke and Ng, 2019), we model text quality by spelling correctness and readability. The former is quantified as absolute and relative counts of three error types from *www.languagetool.org* (hints, unknown words, and others). For the latter, we calculate 10 common readability scores, including Flesch Kincaid Reading Ease, Gunning Fog Index, LIX, and similar.

Evidence On one hand, we capture the evidence given in an argument in terms of the frequency of links. On the other hand, we apply an out-of-the-box argument mining algorithm (Wachsmuth et al., 2016) that classifies each sentence into one of four argumentative unit types: thesis, conclusion, premise, or none.

Subjectivity Since we hypothesized subjectivity to be important, we here combine multiple frequency distributions: (a) singular and plural 1st, 2nd, and 3rd person pronouns, (b) indicators of positivity, negativity, and hedging (e.g., “certainly”) based on Rittman et al. (2004), (c) 83 different emojis, (d) fully lower-case, upper-case, and other words, and (e) character types, such as letter, digit, and whitespace.

5 Experiments

We now report on experiments with the features from Section 4 on the corpus from Section 3. In particular, we systematically study three research questions for the 15 given argument quality dimensions:

Q1. To what extent can each quality dimension be assessed only from an argument's text?

Q2. How dependent is the assessability on the subjective view of the experts?

Q3. How well do the considered features predict argument quality compared to humans?

Experimental Setup We approached all 15 dimensions using each feature type alone, feature ablation (all but one type), and all features respectively. We split the corpus into 16 test sets, one per topic. For each approach and topic, we trained one SVM on the other 15 topics, optimizing its C hyperparameter in 15-fold cross-validation on the training set (tested C range: $10^{-4} \cdot 2^j$ for $7 \leq j \leq 16$). Given that no big outliers exist for the quality score range (1–3), we then computed the *mean absolute error (MAE)*, averaged

		Logical quality				Rhetorical quality						Dialectical quality				
#	**Approach**	**Cog**	**LAc**	**LRe**	**LSu**	**Eff**	**Cla**	**Cre**	**App**	**Emo**	**Arr**	**Rea**	**GAc**	**GRe**	**GSu**	**OvQ**
A_1	Content	0.38	0.42	0.43	0.32	0.34	0.38	0.31	0.36	0.30	0.36	0.39	0.41	0.40	0.24	0.39
A_2	Embedding	0.46	0.44	0.48	0.38	0.39	0.38	0.32	0.36	**0.28**	0.39	0.45	0.47	0.44	0.28	0.45
A_3	Style	0.38	0.41	0.44	0.31	**0.33**	0.39	0.30	0.36	0.29	0.36	0.38	0.40	0.38	0.23	0.38
A_4	Structure	0.45	0.46	0.49	0.38	0.39	0.39	0.34	0.38	0.30	0.41	0.46	0.46	0.44	0.28	0.46
A_5	Length	0.37	**0.40**	0.43	**0.30**	**0.33**	0.38	0.30	0.34	**0.28**	0.35	**0.36**	**0.38**	**0.37**	**0.22**	**0.36**
A_6	Text quality	0.44	0.45	0.48	0.37	0.38	0.38	0.33	0.37	0.31	0.39	0.45	0.44	0.44	0.27	0.44
A_7	Evidence	0.42	0.44	0.46	0.35	0.37	0.40	0.33	0.38	0.30	0.38	0.42	0.45	0.40	0.24	0.42
A_8	Subjectivity	**0.36**	**0.40**	**0.41**	0.31	**0.33**	0.39	0.33	0.35	0.29	0.36	0.37	0.40	0.38	**0.22**	0.37
$A_{\setminus 1}$	w/o Content	0.37	**0.40**	0.43	**0.30**	**0.33**	**0.36**	0.30	**0.33**	0.29	0.35	0.37	0.40	**0.37**	**0.22**	0.37
$A_{\setminus 2}$	w/o Embedding	0.37	**0.40**	0.42	**0.30**	**0.33**	**0.36**	0.30	**0.33**	0.29	0.35	0.37	0.40	**0.37**	**0.22**	0.37
$A_{\setminus 3}$	w/o Style	**0.36**	**0.40**	0.42	**0.30**	**0.33**	**0.36**	0.31	**0.33**	0.29	**0.34**	0.37	0.39	**0.37**	**0.22**	0.37
$A_{\setminus 4}$	w/o Structure	**0.36**	**0.40**	0.42	**0.30**	**0.33**	0.37	**0.29**	**0.33**	0.29	**0.34**	0.37	0.39	**0.37**	**0.22**	0.37
$A_{\setminus 5}$	w/o Length	**0.36**	0.41	0.42	**0.30**	**0.33**	**0.36**	0.30	**0.33**	**0.28**	**0.34**	0.37	0.40	**0.37**	**0.22**	0.37
$A_{\setminus 6}$	w/o Text quality	0.37	**0.40**	0.42	**0.30**	**0.33**	0.37	0.30	0.34	**0.28**	0.35	0.37	0.40	**0.37**	**0.22**	0.37
$A_{\setminus 7}$	w/o Evidence	**0.36**	**0.40**	0.42	**0.30**	**0.33**	**0.36**	0.30	**0.33**	**0.28**	0.35	0.38	0.40	**0.37**	**0.22**	0.37
$A_{\setminus 8}$	w/o Subjectivity	0.37	0.41	0.43	0.31	**0.33**	0.37	0.30	0.34	0.29	0.35	0.37	0.40	**0.37**	**0.22**	0.37
A_{1-8}	All features	‡0.37	‡**0.40**	0.42	‡**0.30**	‡**0.33**	**0.36**	0.30	†**0.33**	0.29	†0.35	†0.37	†0.40	†**0.37**	†**0.22**	‡0.37
B	Baseline	0.44	0.46	0.47	0.39	0.39	0.40	0.33	0.39	0.31	0.40	0.43	0.46	0.43	0.26	0.45

Table 1: Q1. Mean absolute error of each feature type, feature ablation, all features, and the mean baseline for all 15 quality dimensions, averaged over all 16 test sets. The best value in a column is bold. For *all features*, significant improvements over the mean baseline are marked with † ($p < .05$) and ‡ ($p < .01$).

over the 16 MAEs on each test set. This leave-one-topic-out way ensures that no topic information can be exploited in the assessment on the test sets.

To focus on the learning success, we compare the features only to the *mean baseline*, which always predicts the mean score of all arguments in the given training set. For the SVM with all features, we use the 16 single MAEs in a one-tailed independent t-test to test whether improvements over the baseline are significantly better at $p < .05$ (marked † below) and $p < .01$ (‡). The Java code for reproducing the experiments can be accessed here: `http://arguana.com/software`

Quality Assessment (Q1) To provide answers to question Q1, we let all SVMs learn to assess the mean score of the three experts. Table 1 shows the MAE of each feature type (A_i), feature ablation ($A_{\setminus i}$), and all features (A_{1-8}) in comparison to the baseline for each quality dimension. The SVM with all features (A_{1-8}) outperforms the baseline in all cases. Only for four dimensions, the gains are not significant, three of which being rhetorical: clarity (*Cla*), credibility (*Cre*), and emotional appeal (*Emo*). This may be due to their subjective nature, as reflected in limited inter-annotator agreement (Wachsmuth et al., 2017b). The highest MAE reduction is achieved for local sufficiency (0.39 to 0.30), which has also been successfully assessed in previous studies (Stab and Gurevych, 2017). Other clear gains are achieved for overall quality (0.45 to 0.37) and for cogency (0.44 to 0.37). No dimension is really "solved" by the given features, but we conclude that intrinsic argument quality assessment is effective to some extent.

Looking at the features, we see that *content (A_1)* and *embedding (A_2)* perform rather badly, unlike in many NLP tasks. This is somewhat expected, though, due to our leave-one-topic-out setting. While feature ablation leads to the best results for some dimensions (e.g., *Cre* and *Arr*), two feature types dominate the assessment: *Subjectivity (A_8)* alone minimizes the MSE for five dimensions, once being the single best approach (for local relevance, *LRe*). However, *length (A_5)* is even stronger, e.g., being best for reasonableness and overall quality. Albeit quality may require some words, this reveals length bias inherent to the corpus. Such bias was also found in other argument quality corpora (Potash et al., 2017). It questions the validity of the annotated scores, even if A_5 is not needed for many dimensions (see $A_{\setminus 5}$).

Subjectiveness (Q2) For Q2, we learn to assess the score of each single expert and compare our features to the baseline in Table 2. *Mean scores* lead to the lowest MAE, due to their natural tendency towards middle scores. We find clear differences between the experts, reflecting how subjective the assessment is:

		Logical quality				Rhetorical quality						Dialectical quality				
Scores	**Approach**	**Cog**	**LAc**	**LRe**	**LSu**	**Eff**	**Cla**	**Cre**	**App**	**Emo**	**Arr**	**Rea**	**GAc**	**GRe**	**GSu**	**OvQ**
Expert #1	All features	†0.57	0.57	0.52	‡**0.47**	†0.54	0.52	0.41	0.51	0.35	0.54	†0.52	0.56	0.54	‡**0.45**	‡**0.52**
	Baseline	0.64	0.60	0.59	0.57	0.61	0.52	0.48	0.51	0.34	0.55	0.59	0.59	0.56	0.54	0.60
Expert #2	All features	0.49	0.59	0.63	0.34	0.48	0.53	0.37	0.67	0.31	0.49	0.59	0.56	0.58	0.26	‡**0.50**
	Baseline	0.57	0.60	0.67	0.44	0.56	0.53	0.40	0.64	0.28	0.53	0.62	0.56	0.63	0.30	0.62
Expert #3	All features	‡**0.50**	†0.51	†**0.53**	‡**0.47**	†0.38	0.54	†**0.48**	0.33	0.44	0.50	‡**0.46**	0.47	0.53	†0.32	‡**0.47**
	Baseline	0.60	0.57	0.59	0.60	0.45	0.54	0.53	0.37	0.51	0.53	0.55	0.51	0.53	0.40	0.59
Mean score	All features	‡**0.37**	‡**0.40**	0.42	‡**0.30**	‡**0.33**	0.36	0.30	†**0.33**	0.29	†**0.35**	†0.37	†**0.40**	†**0.37**	†0.22	‡ **0.37**
	Baseline	0.44	0.46	0.47	0.39	0.39	0.40	0.33	0.39	0.31	0.40	0.43	0.46	0.43	0.26	0.45

Table 2: Q2. Mean absolute error of the SVM with all features and the mean baseline for all quality dimensions on all test sets, separated for training on the ground-truth scores of expert #1, #2, or #3, or the mean scores (as in Table 1) respectively. The value with the highest significance in each column is bold.

	"Approach"	**Cog**	**LAc**	**LRe**	**LSu**	**Eff**	**Cla**	**Cre**	**App**	**Emo**	**Arr**	**Rea**	**GAc**	**GRe**	**GSu**	**OvQ**
Humans	Expert #1	0.32	0.32	‡**0.23**	0.29	0.32	0.34	‡**0.19**	0.30	**0.18**	**0.27**	†0.22	‡0.27	0.40	0.29	0.27
	Expert #2	0.28	0.38	0.33	0.31	‡0.24	0.31	0.27	0.46	0.25	0.35	0.30	†0.29	0.41	0.24	†0.26
	Expert #3	‡**0.17**	‡**0.22**	‡**0.23**	‡**0.15**	‡**0.16**	**0.26**	0.27	**0.26**	0.27	0.28	‡**0.16**	‡**0.26**	‡**0.22**	‡**0.08**	‡**0.14**
SVM	All features	0.36	0.38	0.42	0.34	0.36	0.33	0.29	0.28	0.22	0.33	0.34	0.40	0.39	0.21	0.36

Table 3: Q3. Mean absolute error of each expert and the SVM with all features, for all quality dimensions on all test sets, based on majority scores. The best value for each dimension is bold. Gray expert values are worse than the features; significant gains over the features are marked with † ($p < .05$) and ‡ ($p < .01$).

Hardly any significant learning success is observed on the scores of *expert #2*, whereas particularly *#3* seems well-assessable. While this may mean that some experts are either more reliable or more influenced by surface text features, it raises the question whether assessing the mean score is the best choice.

Human vs. Machine (Q3) For Q3, finally, we evaluate how much the experts diverge from the majority score as opposed to the all-features SVM. Since the experts could only give integer scores, for fairness we rounded the scores of the SVM before MAE computation. Still, Table 3 reveals that *expert #2* significantly beats our features only on three dimensions (*Eff*, *GAcc*, and *OvQ*) and is even worse on five dimensions (*App*, *Emo*, *Arr*, *GRe*, and *GSu*). So, our features can compete with some humans. *Expert #3*, in contrast, clearly outperforms the SVM with a very low MAE for most dimensions. Together with the results on Q2, it seems that some expert scores are more consistent.

6 Conclusion

In this focused study, we have systematically benchmarked how well argument quality can be assessed computationally on an existing corpus annotated for several quality dimensions. Modeling subjectiveness in terms of sentiment, pronoun usage, and similar seems useful on the debate portal arguments included in the corpus, at least for logical and dialectical dimensions. However, the limited corpus size naturally makes it hard to find more complex features that robustly predict argument quality. In addition, the correlation of quality and length in the corpus limits the generalizability of our findings. This calls for more large-scale and balanced argument quality corpora. First attempts in this direction have been made (Toledo et al., 2019), but the comprehensive view on quality of the corpus used here has no equal so far.

Acknowledgments

We thank Eyke Hüllermeier for suggestions on the experimental setup, Gabriela Molina León for feedback on early drafts, and the anonymous reviewers for their helpful comments.

References

Charu C. Aggarwal and ChengXiang Zhai, 2012. *A Survey of Text Classification Algorithms*, pages 163–222. Springer US, Boston, MA.

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.

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*.

Roxanne El Baff, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2018. Challenge or empower: Revisiting argumentation quality in a news editorial corpus. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 454–464, Brussels, Belgium, October. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the persuasive effect of style in news editorial argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160, Online, July. Association for Computational Linguistics.

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? Choosing the more convincing evidence with a Siamese network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 967–976, Florence, Italy, July. Association for Computational Linguistics.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7805–7813. AAAI.

Ivan Habernal and Iryna Gurevych. 2016a. What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2016b. Which argument is more convincing? Analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599. Association for Computational Linguistics.

Ralph H. Johnson and J. Anthony Blair. 2006. *Logical Self-defense*. Intern. Debate Education Association.

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

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument Strength is in the Eye of the Beholder: Audience Effects in Persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753. Association for Computational Linguistics.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Peter Potash, Robin Bhattacharya, and Anna Rumshisky. 2017. Length, interchangeability, and external knowledge: Observations from predicting argument convincingness. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 342–351, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Martin Potthast, Lukas Gienapp, Florian Euchner, Nick Heilenkötter, Nico Weidmann, Henning Wachsmuth, Benno Stein, and Matthias Hagen. 2019. Argument search: Assessing argument relevance. In *42nd International ACM Conference on Research and Development in Information Retrieval (SIGIR 2019)*, pages 1117–1120. ACM, July.

Robert Rittman, Nina Wacholder, Paul Kantor, Kwong Bor Ng, Tomek StrzaIkowski, and Ying Sun. 2004. Adjectives as indicators of subjectivity in documents. In *Proceedings of the 67th ASIS&T Annual Meeting*, pages 349–359. American Society for Information Science.

Christian Stab and Iryna Gurevych. 2017. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990. Association for Computational Linguistics.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Manfred Stede and Jodi Schneider. 2018. *Argumentation Mining*. Number 40 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment - New datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635. Association for Computational Linguistics.

Frans H. van Eemeren and Rob Grootendorst. 2004. *A Systematic Theory of Argumentation: The Pragma-Dialectical Approach*. Cambridge University Press, Cambridge, UK.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691. The COLING 2016 Organizing Committee.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017a. Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255. Association for Computational Linguistics.

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

Henning Wachsmuth, Benno Stein, and Yamen Ajjour. 2017c. "PageRank" for argument relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1117–1127. Association for Computational Linguistics.

Douglas Walton. 2006. *Fundamentals of Critical Argumentation*. Cambridge University Press.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? Ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200. Association for Computational Linguistics.

Wonsuk Yang, Seungwon Yoon, Ada Carpenter, and Jong Park. 2019. Nonsense!: Quality control via two-step reason selection for annotating local acceptability and related attributes in news editorials. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2954–2963. Association for Computational Linguistics.