AQUAPLANE: The Argument Quality Explainer App

Sebastian Britner
sebastianbritner@gmail.com
Trier University
Trier, Germany

Lorik Dumani
dumani@uni-trier.de
Trier University
Trier, Germany

Ralf Schenkel
schenkel@uni-trier.de
Trier University
Trier, Germany

ABSTRACT

In computational argumentation, so-called quality dimensions such as coherence or rhetoric are often used for ranking arguments. However, the literature often only predicts which argument is more persuasive, but not why this is the case. In this paper, we introduce Aquaplane, a transparent and easy-to-extend application that not only decides for a pair of arguments which one is more convincing with respect to a statement, but also provides an explanation.

CCS CONCEPTS

• **Information systems** → *Information systems applications*; *Web applications*; *Information retrieval query processing*; *Retrieval models and ranking*.

KEYWORDS

argumentation, argument quality, explanations

ACM Reference Format:
Sebastian Britner, Lorik Dumani, and Ralf Schenkel. 2023. AQUAPLANE: The Argument Quality Explainer App. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3583780.3614733

1 INTRODUCTION

Argumentation is an essential part of human communication when there are divergent opinions or conflicts of interest [9]. People argue, among others, in social media, newspaper articles, and political speeches. The goal of argumentation is to persuade an audience, reach agreements, resolve disputes, portray justifications, and find decisions [36]. In the field of computational argumentation (CA), an *argument* is defined as a *claim* that is supported or opposed by at least one *premise* [45]. While the claim portrays a controversial standpoint for which the speaker wants an audience to either increase or decrease its acceptance, a premise serves as evidence or clue to do so. The polarity from a premise to the claim, i.e., whether it is supporting or rejecting, is defined as *stance*. An example for a (controversial) claim is "*tv is better than books*", a supporting premise to this claim is "*Books and newspapers can't give you emergency warnings*", an opposing premise is "*watching tv has a negative effect on mental health*".

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3614733

The research area of CA includes tasks such as extracting arguments from natural language texts (*argument mining*) [22], classifying arguments into their viewpoints (*stance prediction*) [34], retrieving and ranking arguments to a query (*argument retrieval*) [42], or generating new arguments (argument generation) [1]. In this paper, we address a subtask of argument retrieval by considering the *quality of arguments* for ranking [6], as it has a significant impact on whether an argument can achieve its goals [41]. However, the literature often only predicts ***which*** argument is more persuasive or of higher quality, but not ***why*** this is the case. Explaining such choices is essential because the effect of an argument on a person may differ due to distinct values and their weighting [2]. For instance, a person might regard an argument as good if convinced by the truth of its premise, while another person might be convinced by a persuasive language. These different qualities are called *argument quality dimensions* [41]. This subjectivity in perceiving the effect of arguments implies the necessity to additionally show explanations for assigning a higher quality to an argument. A positive side effect is that it establishes more trust to decisions made by an automated system.

In this paper we address explainable argument quality. More precisely, given a pair of arguments with the same stance regarding a controversial claim, our goal is not only to decide which argument is more convincing overall and in several argument quality dimensions, but also to automatically explain and justify this decision. We present Aquaplane, the **A**rgument **QUA**lity ex**PLAiNE**r, a transparent, modular, extensible, and interactive system[1]. Given a claim and two premises with the same stance, it compares them in 15 quality dimensions and explains its decisions. Users can interactively explore the customized explanations to understand the decisions for each dimension.

2 RELATED WORK

Habernal and Gurevych [15] present the dataset *UKPConvArg1* which consists of 16k pairs of arguments each with the same stance on the same topic collected from debate portals. They introduce a relative approach to evaluate their persuasiveness by picking the more convincing argument. The methods are promising but it turns out to be complex to derive the reasons for the decision from these models. Gleize et al. [10] take evidence from Wikipedia into account when assessing persuasiveness. They propose a Siamese neural network to solve the task which outperforms the aforementioned approach [15]. Potash et al. [31] and Gleize et al. [10], among others, find a length bias in the UKP datasets, causing methods that use text length to determine persuasiveness to produce results similar to deep learning models. Toledo et al. [37] provide the state-of-the-art approach to determine the more persuasive

[1] Code and demo video: https://github.com/recap-utr/Aquaplane.

argument through binary text classification with BERT [4]. For this purpose, they conduct an annotation study on 6.3k arguments collected by using *Speech by Crowd*, a service developed by IBM to support the collection of arguments. Among others, they prevent length bias by lower differences in text lengths and by limiting text length. However, this does not capture deeper reasoning and more complex argumentation. Further, it is also not robust for use in the real world, where texts are not curated but noisy.

Habernal and Gurevych [14] use the natural language justifications for the decisions which are captured next to the labels in the dataset *UKPConvArg1* to evaluate the qualitative properties in each argument pair. Their corpus *UKPConvArg2* consists of 9,111 argument pairs annotated with 17 categories targeting different aspects such as information content, subjectivity or comprehensibility. The evaluation showed that a fine-grained analysis of the persuasiveness of arguments requires further investigation. Wachsmuth et al. [41] analyzed various argument quality dimensions from the literature and divide the overall quality into logical, rhetorical, and dialectical quality, which in turn can be divided into sub-dimensions. They provide a taxonomy comprising 15 argument quality dimensions and *Dagstuhl-15512 ArgQuality Corpus*, a dataset of arguments annotated with respect to the different quality dimensions. Their work was a cornerstone for a lot of further works [5, 12, 32, 48]. Based on this corpus, Wachsmuth and Werner [43] examine which linguistic features of a text can be used to evaluate the different dimensions of argument quality. They establish eight features quantified using various aspects such as spelling errors, use of personal pronouns, length of sentences and words, and types of argument units. They achieved moderate, yet significant, success for scoring most dimensions. However, due to the small size of the dataset, it was not possible to identify additional and more complex features. El Baff et al. [8] investigate how the style of a news article influences persuasiveness, showing that stylistic features have a greater influence on predicting persuasiveness among certain readers than content features. Persing and Ng [30] measure how unconvincing an argument is while also examining why an argument is unconvincing. They define five types of errors and annotate a corpus of arguments from debates with their persuasiveness and thus which errors the author committed. It remains an open question whether the error types are specific enough to help authors identify errors concretely and thus make arguments more persuasive.

3 ARGUMENT QUALITY DIMENSIONS

We now review the 15 logical, rhetorical, and dialectical quality dimensions for arguments from Wachsmuth et al. [41] for which we implemented methods for measurement. The logical quality considers if the reasons given for an argument are reasonable and comprehensible. The rhetorical perspective evaluates how effectively an argument is presented, and the dialectical perspective whether objections are adequately refuted by the argument. They distinguish between higher-level dimensions and sub-dimensions and provide definitions for them:

***Cogency* (*Co*)** (refers to the *logical* quality): The premises of an argument are acceptable, relevant to the conclusion, and sufficient to draw it. • *Local Acceptability* (*LA*): The premise of an argument is rationally worth believing to be true. • *Local Relevance* (*LR*): The premise of an argument contributes to the acceptance or rejection of the conclusion of the argument. • *Local Sufficiency* (*LS*): The premises of an argument are sufficient to draw the conclusion.

***Effectiveness* (*Ef*)** (refers to the *rhetorical* quality): The argument convinces the target audience of the author's stance on a particular issue. • *Credibility* (*Cr*): The argument is conveyed in a way that makes the author seem credible. • *Emotional Appeal* (*Em*): The emotions generated by the argument make the target audience more open to the author's arguments. • *Clarity* (*Cl*): The argument uses correct and clear language, avoids unnecessary complexity, and does not stray from the topic. • *Appropriateness* (*Ap*): The language used in the argument supports the emergence of credibility and emotion and is appropriate to the topic. • *Arrangement* (*Ar*): The topic, arguments, and conclusion are placed in the argument in a proper order.

***Reasonableness* (*Re*)** (refers to the *dialectical* quality): The argument makes a sufficient contribution to the solution of the problem and is accepted by the target audience. • *Global Acceptability* (*GA*): The target audience accepts both the consideration of the arguments given and the way they are portrayed. • *Global Relevance* (*GR*): The argument cites information and arguments that lead to a final conclusion and thereby contribute to problem solving. • *Global Sufficiency* (*GS*): The argument adequately refutes expected counterarguments.

The *Overall Quality* is the general assessment of quality. In this paper, it is considered as a function of the other dimensions.

4 MAPPING METHODS TO QUALITIES

We now present the methods we use to (i) measure and determine the argument quality dimensions from Section 3 and (ii) explain the decision which argument is better. Note that our mapping is based on theoretical assumptions which we justify below. Note that we kept the mapping of the methods as well as adding or removing them to the dimensions flexible in the code.

Implemented Methods. **Profanity:** Profanity refers to the use of unacceptable, insulting, or offensive language in the form of cursing [24]. We employ the blacklist by Parker [26] to detect it, setting the profanity of an argument arg as $\frac{\text{number of profane words in } arg}{\text{number of words in } arg}$. It has a negative impact on *Cl* and *Ap*, since this inappropriate language makes the author look unprofessional, immature, and thus untrustworthy.

Fact-Checking: To prevent negative effects of misinformation it is necessary to check the correctness and reliability of information with *fact-checking*. Automated fact-checking systems [17] often divide the task into three stages [13]: identifying claims to be verified (*check-worthiness*), collecting relevant information, and assigning truthfulness. We use the ClaimBuster API [19] to determine check-worthiness and check the truthfulness through the Google FactCheck Claim Search API [11]. We then determine the similarity of the yielded claims using SBERT and cosine similarity to the clause part, and only proceeded with the most similar one. We trained a RoBERTa model [23] on the MNLI [47] dataset to detect the stance and invert the ratings if necessary. We map these cosine values to *LA* and *GA* because false claims in an argument lead to less acceptance.

Spell Check: Spell checking is necessary to guarantee correct language usage. We follow a rule-based approach [25] to detect spelling errors. For an argument, the number of misspelled and unknown words is related to the argument length in words. Spelling errors have an indirect influence on *Cr* as many spelling errors can make an author look unprofessional, and on *Cl* as arguments with fewer spelling errors are more readable and lead to fewer comprehension problems.

Stylometry: Stylometry refers to the analysis of linguistic features of a natural language text to capture and characterize an author's writing style [20]. We use a subset of the stylometric features implemented in StyleExplorer [38] and map these to the *Cl* because a complex sentence structure and vocabulary can lead to an argument not being understood or even misunderstood.

Search Engine for SimpleWiki: Wikipedia serves as a modern online lexicon for general knowledge. We indexed the *simpleWiki* [46] dump (417,965 entries) with Apache Lucene [35] (version 9.4.1), applying BM25F [28]. We use this to get the most relevant SimpleWiki article to an argument and claim with its BM25F score. Since *simpleWiki* [46] provides general knowledge and the query uses claim and argument, we assume that a higher BM25F score means that the argument contains information that is more generally relevant. Thus, the method influences *LR* and *GR*.

Search Engine for debate-org: In debate forums, people argue on controversial topics in order to convince opponents to a particular standpoint. We use the DDO dataset by Durmus and Cardie [7] consisting of 51,594 debates to estimate the relevance of arguments, creating a search engine similar to the one for SimpleWiki. Each entry in DDO includes the claim, as well as pro and con arguments. We infer the relevance of an argument which we use as query on the basis of the highest BM25F score. Since a search query consists of an claim and an argument, we conclude from a high BM25F score that an argument is generally more relevant to solve problems. Therefore, this method is used to determine *GR*.

URL Sources: Arguments may include sources placed to support claims, often providing sources in the form of URLs. To detect URLs within arguments, we use a regular expression of Perini [29]. By adding sources, both argument and author may appear more credible. Thus, the number of sources employed serves gauging *Cr*.

Excessive Punctuation: We define excessive punctuation to be a sequence of three or more punctuation marks. We assume that it has a negative impact on argument quality: an author repeatingly using exclamation points or question marks may appear angrier or guided by emotion. Thus, anger makes a person appear inexpertly and is judged less appropriate [39]. Therefore, we assume that excessive punctuation has a negative influence on *Cr*, *Em*, and *Ap*.

All-Caps-Words: *All-Caps-Words* are words composed of capital letters only, mostly used for emphasizing, or to express emotions. The use of all-caps words could be construed as shouting in the context of social media [18], but also indicate emotional states such as anger, excitement, or joy. We hypothesize that all-caps words have a negative impact on the dimension of *Em* and that shouting has a negative influence on *Cr* and *Ap*.

Dramatic Language: We define dramatic language as descriptive and figurative with metaphors, exaggerations, and other rhetorical stylistic devices. We adopt a simple list-based approach to recognize it, using the adverb lists provided by Rashkin et al. [33]. We determine the dramatic nature of the language as the fraction of these adverbs among all words of an argument. We suppose that dramatic language has a positive influence on the dimension *Em*.

Ad-hominem-arguments: *ad-hominem*-arguments attack individuals based on their characteristics or circumstances, rather than making reference to a counterargument [44]. They represent a fallacy within an argument. Based on the strong results obtained by Patel et al. [27], we finetune ALBERT [21] on the dataset of Habernal et al. [16] for the binary detection of ad hominem arguments. We assign this method to the dimension *GA* since Wachsmuth et al. [40] show that the justification that an argument is attacking or offensive correlates strongly with *GA*. This goes along with the view that personal attacks are generally unacceptable.

Determining the more convincing argument. Given a pair of arguments, we apply all methods to obtain scores for both. We declare the argument with a higher score for a method to be the comparatively better one for it. There are three outcomes for each dimension and pair: `0`: no decision possible, 1: decision for argument 1, and 2: decision for argument 2. While some methods return binary scores (e.g. ad-hominem), some even return multiple scores (e.g. stylometry). The final decision then results from the Boyer-Moore Majority Vote algorithm [3]. To obtain the scores for subdimensions and dimensions, the mapped methods and the hierarchical structure are applied by majority decision. For example, the decision for the subdimension *Em* results from the decisions of *excessive punctuation* and *dramatic language*. This makes the decision-making process transparent. Since there are few methods for mapping so far and no methods could be assigned to the *LS*, *Ar*, and *GS*, ambiguous decisions (`0`) occur frequently. To prevent a strong impact, these were removed from majority decisions. In future work, we expect Aquaplane to be augmented by more methods.

Generation of explanations. We generate static explanations corresponding to predefined templates for all decisions for all 15 dimensions. While the decision making process is bottom-up, i.e. from the decisions of the methods to the overall decision, the explanations are presented top-down. The explanations are presented in stages, intended to achieve the interactive aspect of explanations. The overall decision on *Overall Quality* is summarized in a short statement, e.g. "*Argument 1 is more convincing than argument 2*". This overall decision can be explained by the decisions on the next lower dimensions, e.g. "*Argument 1 is more (cogent or effective or reasonable) than argument 2", because of its [dimension]*". Likewise, their subdimensions are explained by "*Argument 1 has a higher [subdimension] than argument 2*". Lastly, for each method, the values computed and the information returned are presented by applying customized explanations (as they include pieces of arguments inserted in the tool) to provide more understanding. For example for the dimension *Cr* and the method URL sources an explanation could be "*Argument 1 generally gives more sources. The following sources were provides: http://example.org, http://anotherexample.org*".

5 EVALUATION

We now examine how well the qualities presented in Section 3 can be determined by the methods detailled in Section 4.

Table 1: Results for the evaluation dataset

Quality dimension		Acc.	BL	Macro-F1	BL
Co	**Cogency**	.37	.45	.32	.21
LA	Local Acceptability	.42	.42	.20	.20
LR	Local Relevance	.42	.41	.35	.19
LS	Local Sufficiency	-	.53	-	.23
Ef	**Effectiveness**	.25	.53	.25	.23
Cr	Credibility	.36	.53	.36	.23
Em	Emotional Appeal	.33	.65	.31	.26
Cl	Clarity	.33	.45	.32	.21
Ap	Appropriateness	.38	.43	.36	.20
Ar	Arrangement	-	.44	-	.20
Re	**Reasonableness**	.42	.45	.42	.21
GA	Global Acceptability	.42	.38	.31	.19
GR	Global Relevance	.46	.36	.45	.18
GS	Global Sufficiency	-	.68	-	.27
Ov	**Overall Quality**	.40	.44	.38	.20
Mc	**More Convincing**	.64	.51	.45	.34

Dataset. We derive an evaluation dataset from the datasets *UKPConvArg1* [15] and *Dagstuhl-15512-ArgQuality* [41]. *UKPConvArg1* contains argument pairs from debate portals with the same viewpoint on 16 topics. We use the version *UKPConvArg1Strict* where argument pairs with equal persuasiveness were removed. The *Dagstuhl-15512-ArgQuality* corpus contains assessments of 320 arguments from the *UKPConvArg1* corpus on the 15 argument quality dimensions. For the assessments, three experts assigned a value to each argument regarding the different dimensions on the scale from 1 (Low) to 3 (High). We take the median of these three ratings for each argument for each dimension. Further, we only use argument pairs from *UKPConvArg1Strict* if both arguments are in the corpus *Dagstuhl-15512-ArgQuality* which holds for 985 pairs. For these, we compare the scores on each dimension and derive a decision value that numerically identifies the argument with the higher score. For all 985 instances, we now determine the more convincing argument with Aquaplane and compare them with the labels of the evaluation dataset. We create a baseline for each dimension, where we always take the decision that occurs most frequently in a dimension.

Results. Table 1 shows the calculated accuracies and macro F_1-scores to the decisions of each dimension together with the baselines (BL). Only for a few dimensions are the accuracy values above the baselines. Even though the accuracy values and F_1-scores are quite low, a good tendency can be seen for some dimensions like *GR*, which indicates that the assigned methods have a positive influence. In general, however, the accuracy values and F_1 scores are not satisfactory. In a manual investigation, we found that some of the methods are not mature and can generate errors.

Determination of the Overall Quality. We evaluated the extent to which the more convincing argument can be determined by a majority decision from the dimensions. Specifically, for each of the dimensions *Co*, *Ef*, and *Re*, we tested whether their decision value follows from the majority decision of the respective subdimensions. Further, we tested whether the decision on the *Overall Quality* or the *More-Convincing* label taken from the *UKPConvArg1* dataset follows from the majority decision of the *Co*, *Ef*, and *Re* dimensions.

The dimensions *Co* and *Re* in many cases infer their decision by a majority vote from their assigned subdimensions. Thus, these dimensions are shown to be well, but not fully, represented by their subdimensions. In contrast, *Ef* seems to be much more difficult

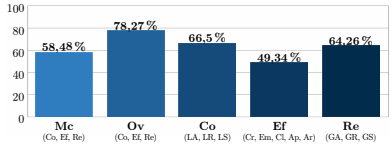

Figure 1: Frequency of agreement of the decision on a dimension by majority voting.

to determine by a majority decision of its subdimensions. This could indicate that the subdimensions encompass more than is captured by the *Ef* dimension. This could also account for another reason for the very low accuracy and F_1 score for *Ef* in the Table 1. The *Overall Quality* can often be derived from the *Co*, *Ef*, and *Re* dimensions by majority vote at 0.7827. This is also consistent with the correlation that Wachsmuth et al. [41] measured. When all dimensions are added to the decision of *Overall Quality* as a test, the frequency reduces to 0.603, which shows that deriving *Overall Quality* from the three previously mentioned dimensions is a good choice. Figure 1 illustrates this.

6 APPLICATION

The application provides transparency to the decisions as it presents its explanations. In addition, researchers as well as interested users can interactively navigate through the generated explanations to gain understanding of the decision.

A user can enter two arguments together with the claim they refer to. Alternatively it would also be possible to upload a CSV file to enable calculations for multiple inputs. By clicking a button, for both arguments each method to compute the argument quality will be processed, then the qualities will be compared, and explanations of the decisions presented. Users can then interactively navigate through the explanations to gain a deeper understanding of the decision if needed. Interaction happens by clicking on specific terms (highlighted in color) within the explanation text, e.g. clicking on the term "*clarity*" in the text "Argument 1 has a higher *Cl* than argument 2.", which explains in a detailed view why the argument has a higher *Cl*. The results can be downloaded in a JSON file along with all the information used in the argument quality comparison. Figure 2 shows the application.

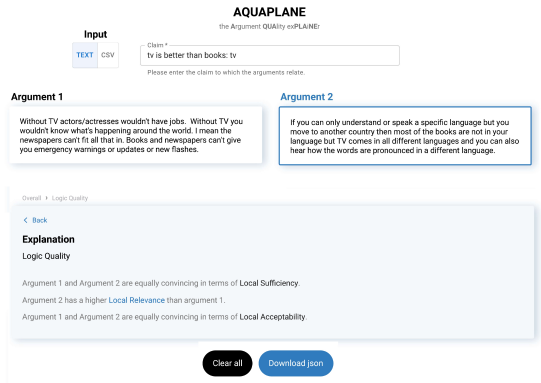

Figure 2: GUI of Aquaplane.

REFERENCES

[1] Khalid Al Khatib, Lukas Trautner, Henning Wachsmuth, Yufang Hou, and Benno Stein. 2021. Employing argumentation knowledge graphs for neural argument generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 4744–4754.

[2] Trevor Bench-Capon. 2021. Audiences and Argument Strength.. In *ArgStrength 2021* (11/10/2021 - 13/10/2021). Hagen (online).

[3] Robert S Boyer and J Strother Moore. 1991. MJRTY: A Fast Majority Vote Algorithm. *Automated reasoning: essays in honor of Woody Bledsoe* 1 (1991), 105–117.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[5] Lorik Dumani and Ralf Schenkel. 2020. Quality-Aware Ranking of Arguments. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (2020). https://api.semanticscholar.org/CorpusID:224281338

[6] Lorik Dumani, Tobias Wiesenfeldt, and Ralf Schenkel. 2021. Fine and Coarse Granular Argument Classification before Clustering. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, Gianluca Demartini, Guido Zuccon, J. Shane Culpepper, Zi Huang, and Hanghang Tong (Eds.). ACM, 422–432. https://doi.org/10.1145/3459637.3482431

[7] Esin Durmus and Claire Cardie. 2018. Exploring the Role of Prior Beliefs for Argument Persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1035–1045. https://doi.org/10.18653/v1/N18-1094

[8] 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*. Association for Computational Linguistics, Online, 3154–3160. https://doi.org/10.18653/v1/2020.acl-main.287

[9] Austin J Freeley and David L Steinberg. 2008. *Argumentation and Debate* (12 ed.). Cengage Learning.

[10] 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*. Association for Computational Linguistics, Florence, Italy, 967–976. https://doi.org/10.18653/v1/P19-1093

[11] Google. 2018. Google FactCheck Claim Search API. https://factchecktools.googleapis.com/v1alpha1/claims:search.

[12] Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2019. A Large-scale Dataset for Argument Quality Ranking: Construction and Analysis. *ArXiv* abs/1911.11408 (2019). https://api.semanticscholar.org/CorpusID:208291067

[13] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics* 10 (2022), 178–206.

[14] Ivan Habernal and Iryna Gurevych. 2016. 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*. Association for Computational Linguistics, Austin, Texas, 1214–1223. https://doi.org/10.18653/v1/D16-1129

[15] Ivan Habernal and Iryna Gurevych. 2016. 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)*. Association for Computational Linguistics, Berlin, Germany, 1589–1599. https://doi.org/10.18653/v1/P16-1150

[16] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before Name-Calling: Dynamics and Triggers of Ad Hominem Fallacies in Web Argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 386–396. https://doi.org/10.18653/v1/N18-1036

[17] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. 2017. Claimbuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment* 10, 12 (2017), 1945–1948.

[18] Maria Heath. 2021. No need to yell: A prosodic analysis of writing in all caps. *University of Pennsylvania Working Papers in Linguistics* 27, 1 (2021), 10.

[19] IDIR Lab. 2017. ClaimBuster API. https://idir.uta.edu/claimbuster/api/v2/.

[20] Ksenia Lagutina, Nadezhda Lagutina, Elena Boychuk, Inna Vorontsova, Elena Shliakhtina, Olga Belyaeva, Ilya Paramonov, and PG Demidov. 2019. A survey on stylometric text features. In *2019 25th Conference of Open Innovations Association (FRUCT)*. IEEE, 184–195.

[21] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).

[22] John Lawrence and Chris Reed. 2019. Argument Mining: A Survey. *Comput. Linguistics* 45, 4 (2019), 765–818. https://doi.org/10.1162/coli_a_00364

[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[24] Shervin Malmasi and Marcos Zampieri. 2017. Detecting Hate Speech in Social Media. *CoRR* abs/1712.06427 (2017). arXiv:1712.06427 http://arxiv.org/abs/1712.06427

[25] Marcin Miłkowski. 2010. Developing an open-source, rule-based proofreading tool. *Software: Practice and Experience* 40, 7 (2010), 543–566.

[26] James Parker. 2022. Full List Of Bad Words Banned By Google. https://www.freewebheaders.com/full-list-of-bad-words-banned-by-google/.

[27] Utkarsh Patel, Animesh Mukherjee, and Mainack Mondal. 2022. "Dummy Grandpa, do you know anything?": Identifying and Characterizing Ad hominem Fallacy Usage in the Wild. *arXiv preprint arXiv:2209.02062* (2022).

[28] Joaquín Pérez-Iglesias, José R Pérez-Agüera, Víctor Fresno, and Yuval Z Feinstein. 2009. Integrating the probabilistic models BM25/BM25F into Lucene. *arXiv preprint arXiv:0911.5046* (2009).

[29] Diego Perini. 2018. Regular Expression for URL validation. GitHub Gist. https://gist.github.com/dperini/729294

[30] Isaac Persing and Vincent Ng. 2017. Why Can't You Convince Me? Modeling Weaknesses in Unpersuasive Arguments. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 4082–4088. https://doi.org/10.24963/ijcai.2017/570

[31] 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, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, Greg Kondrak and Taro Watanabe (Eds.). Asian Federation of Natural Language Processing, 342–351. https://aclanthology.org/I17-1035/

[32] Niklas Rach, Yuki Matsuda, Johannes Daxenberger, Stefan Ultes, K. Yasumoto, and Wolfgang Minker. 2020. Evaluation of Argument Search Approaches in the Context of Argumentative Dialogue Systems. In *International Conference on Language Resources and Evaluation*. https://api.semanticscholar.org/CorpusID:218973773

[33] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*. 2931–2937.

[34] Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance detection benchmark: How robust is your stance detection? *KI-Künstliche Intelligenz* (2021), 1–13.

[35] The Apache Software Foundation. 1999. Apache Lucene (v9.4.1). https://lucene.apache.org/.

[36] Christopher W. Tindale. 2007. *Fallacies and Argument Appraisal*. Cambridge University Press. https://doi.org/10.1017/CBO9780511806544

[37] 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. *CoRR* abs/1909.01007 (2019). arXiv:1909.01007 http://arxiv.org/abs/1909.01007

[38] Michael Tschuggnall, Thibault Gerrier, and Günther Specht. 2019. StyleExplorer: A Toolkit for Textual Writing Style Visualization. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part II 41*. Springer, 220–224.

[39] Jonathan Van't Riet, Gabi Schaap, and Mariska Kleemans. 2018. Fret not thyself: The persuasive effect of anger expression and the role of perceived appropriateness. *Motivation and Emotion* 42 (2018), 103–117.

[40] Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017. Argumentation Quality Assessment: Theory vs. Practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, 250–255. https://doi.org/10.18653/v1/P17-2039

[41] Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational Argumentation Quality Assessment in Natural Language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, 176–187. https://aclanthology.org/E17-1017

[42] Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. Building an Argument Search Engine for the Web. In *Proceedings of*

the 4th Workshop on Argument Mining. Association for Computational Linguistics, Copenhagen, Denmark, 49–59. https://doi.org/10.18653/v1/W17-5106

[43] Henning Wachsmuth and Till Werner. 2020. Intrinsic Quality Assessment of Arguments. *CoRR* abs/2010.12473 (2020). arXiv:2010.12473 https://arxiv.org/abs/2010.12473

[44] Douglas Walton. 1998. *Ad hominem arguments*. University of Alabama Press.

[45] Douglas Walton, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge and New York: Cambridge University Press.

[46] Wikimedia Foundation, Inc. 2022. simpleWiki. https://dumps.wikimedia.org/simplewiki/20221101/

[47] Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. *CoRR* abs/1704.05426 (2017). arXiv:1704.05426 http://arxiv.org/abs/1704.05426

[48] Tim Ziegenbein, Shahbaz Syed, Felix Lange, Martin Potthast, and Henning Wachsmuth. 2023. Modeling Appropriate Language in Argumentation. *ArXiv* abs/2305.14935 (2023). https://api.semanticscholar.org/CorpusID:258865955