Let's discuss! Quality Dimensions and Annotated Datasets for Computational Argument Quality Assessment

Rositsa V Ivanova and **Thomas Huber** and **Christina Niklaus**
University of St. Gallen, Switzerland
{rositsa.ivanova, thomas.huber, christina.niklaus} @unisg.ch

Abstract

Research in the computational assessment of Argumentation Quality has gained popularity over the last ten years. Various quality dimensions have been explored through the creation of domain-specific datasets and assessment methods. We survey the related literature (211 publications and 32 datasets), while addressing potential overlaps and blurry boundaries to related domains. This paper provides a representative overview of the state of the art in Computational Argument Quality Assessment with a focus on annotated datasets. The aim of the survey is to identify research gaps and to aid future discussions and work in the domain.

1 Introduction

Argumentation is both a key competence and an important cultural technique in democratic societies (Hess, 2009). It serves as a fundamental device for expressing beliefs, perspectives, or justifications around a specific claim. The primary goal of argumentation is to strengthen or weaken the acceptability of a position by presenting supporting or opposing evidence (Eemeren et al., 1996).

In recent years, the field of Argument Mining (AM), i.e., the extraction of arguments from natural language text, has made significant progress (e.g., Trautmann et al., 2020; Morio et al., 2022; Galassi et al., 2023). However, the automatic quality assessment of argumentation (i.e., Argumentation Quality - AQ) is still an open challenge, since defining and measuring the quality of an argument is complex and multifaceted, involving aspects such as logical soundness, persuasiveness, and dialectical reasonableness (e.g., Ng et al., 2020; Gretz et al., 2020; Alhamzeh, 2023). The insights and approaches from AM and AQ have been applied to various research directions. Argument Search (AS) (e.g., Stab et al., 2018; Nilles et al., 2021) makes use of the quality of arguments as an additional aid for the ranking of mined arguments (Wachsmuth et al., 2017b). Argument Improvement (AImp) focuses on the analysis of the quality of an argument in terms of their improvement from a previous version (Zhang et al., 2016a, 2017; Afrin and Litman, 2018; Skitalinskaya et al., 2021).

This paper surveys the literature on Computational Argument Quality Assessment. As our main contributions (1) we summarize the development of the field and the quality dimensions overtime, (2) we provide a detailed analysis of the existing annotated datasets in regards to their size, language, quality dimensions, annotation scales, annotation process, and availability, and (3) we identify significant research gaps and propose concrete research directions to guide future work in this domain.

2 Methodology

We defined the scope of our survey as the scope of the view of computer science on AQ and selected the Digital Bibliography and Library Project (DBLP)[1] as our initial source of publication on the topic. We queried the bibliography using the search term "argument quality" and collected a total of 80 items (i.e., journal articles, conference and workshop papers, informal and other publications) dated up until the end of March 2024[2]. The search function treats the words "argument" and "quality" as individual substrings and matches them to any of the collected metadata (e.g., author, title, venue, type, access, volume, year, URL), thus implicitly expanding the scope of the found publications beyond the strict domain of AQ. We excluded duplicated items (e.g., preprints), unavailable entries or those not referring to actual publications, and work not directly concerning the AQ domain (see Figure 1). Further, we expanded the scope of our

[1] https://dblp.org

[2] We manually searched and included publications from conference proceedings from EACL 2024, which had not yet been published on DBLP. This resulted in 4 additional publications, which are included in the Snowballing count.

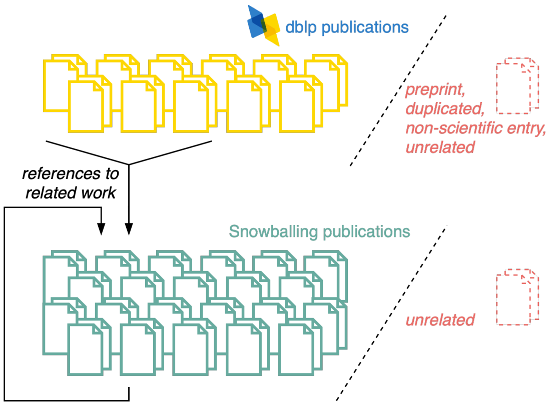

Figure 1: Overview of the applied collection method for publications in the domain.

study by applying Snowball sampling (Goodman, 1961) in an iterative manner. Here, in a first step, we extracted references to prior related work. We then filtered out any publications which were not directly related to the domains Argumentation Quality, Argument Mining and Argument Improvement, such as ones from the philosophical field. At this point we refrained from narrowing down the discovered further references only to the AQ domain in order to achieve a more representative overview. In a second step, we extracted references from the newly collected publications and proceeded with the same filtering approach from the first step. This approach was iterated for all publications until no new relevant references were discovered. This resulted in a total number of 211 publications (see Appendix A for a complete list). Lastly, for the further analysis presented in our paper, we focused particularly on the publications that are relevant to the computational assessment of AQ.

Figure 2 offers an overview of the publications related to the AQ domain over the last 20 years. Starting from 2003, the blue dashed line describes all publications found in DBLP, the solid green line – all publications found in DBLP and the ones collected via Snowballing, and the red bars indicate the number of datasets released per year. Despite the fluctuations in the numbers over the years, we believe that the plots indicate a still rising interest in the field of AQ.

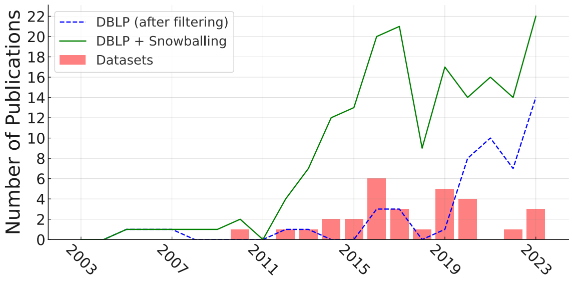

Figure 2: Number of publications (blue dashed line - DBLP (after filtering), green solid line - DBLP and snowballing) and number of new datasets (red bars) throughout the last 20 years (2003 to 2023).

We consider this collection to be representative of the state of the art in Computational Argument Quality Assessment to the best of our knowledge but make no claim to completeness. The following section provides an overview of previous research by examining the proposed quality dimensions.

3 Dimensions for Computational Argumentation Quality Assessment

The applications in the NLP community have been explored in various context. In the field of AQ, Persing and Ng first put their main focus on automated essay scoring. They began refining the holistic scoring schemes used by scoring engines at that time by addressing more specific quality dimensions - starting with organization (Persing et al., 2010), thesis clarity (Persing and Ng, 2014), prompt adherence (Persing and Ng, 2014), and argument strength (Persing and Ng, 2015). The latter shifts the focus towards argumentative essays, while the others explore essays in general. Nevertheless, slowly but surely the analysis of texts went beyond the assessment of structural aspects (something that still remains the strong focus of AM) and began looking for other means to measure the quality of arguments.

As more authors took interest in argumentative texts, they began deriving those from reviews, forum posts, etc. Wachsmuth et al. (2014), for instance, explored sentiment, while Braunstain et al. (2016) took interest in the level of support present in recommendations or opinions. With the increasing number of interactions taking place online, the interest in these interactions remained and a number of publications looked at dimensions such as persuasiveness (e.g., Tan et al., 2016; Persing and Ng, 2017b) and convincingness (e.g., Habernal and Gurevych, 2016a) as means to assess the quality of arguments. In contrast to prior work on persuasion, El Baff et al. (2018) and Durmus et al. (2019)'s work accounts for external factors such as the prior belief of the readers and aims to incorporate and explain subjectivity in the assessment. A particular

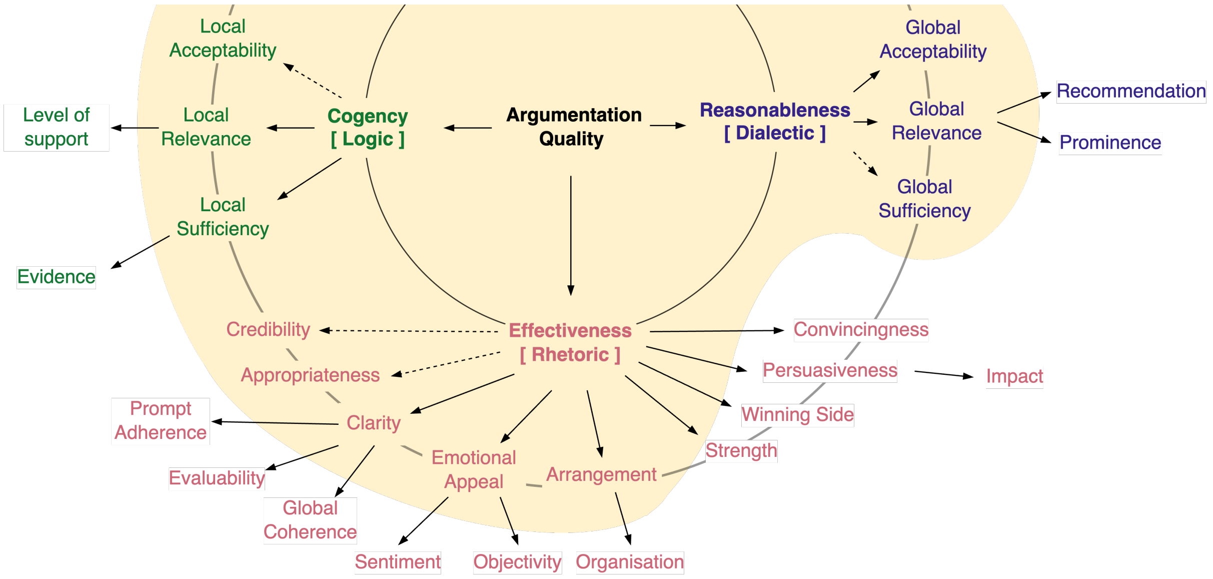

Figure 3: Overview of quality dimensions for computational argument quality assessment discovered in the surveyed literature. *Note: The figure is an extension of a taxonomy proposed in Wachsmuth et al., 2017a (Figure 1), which is highlighted by the yellow background.*

interest on convincingness and recommendedness is noticeable in work by researchers at IBM (e.g., Gretz et al., 2020), who actively collect the raw data (i.e., unlabeled data) through crowdsourcing, instead of extracting it from forums or similar.

In parallel to these developments, which were rather based on intuitive understandings of the targeted quality dimensions, Stab and Gurevych (2017) and Wachsmuth et al. (2017a) based their work on theoretical frameworks previously proposed in philosophy. Following a detailed research on quality dimensions for computational assessment and several major theories for argumentation, Wachsmuth et al. (2017a) created a new taxonomy. Their concept revolves around three high-level quality dimensions defined by Blair (2011) - *logical*, *rhetorical*, and *dialectical* - and adds one more layer of sub-dimensions to each of them (e.g., logical: acceptability, relevance, and sufficiency). With a brief delay, this taxonomy became a significant reference and found applications for the creation of new datasets. For instance Ng et al. (2020) aimed at a domain-diverse set of texts, while Alhamzeh (2023) assessed conference calls in the financial domain. As such the work in the field of AQ once again shifted from an assessment (mostly) of a single quality dimension to the consideration (and thus annotation) of multiple at the same time.

Enhanced Information Utilization for AQ Assessment. Prior work (Hulpus et al., 2019; Lauscher et al., 2022; Plenz et al., 2023b) has argued that it is beneficial to add context to argument via relevant external knowledge in order to better assess its quality. On the contrary, others (e.g. Swanson et al., 2015; Wachsmuth and Werner, 2020; Plenz et al., 2023a) extract syntactic features (e.g., sentence length, vocabulary richness) from the provided text to aid the quality assessment. Further, Sun et al. (2021) take a look at the syntax of arguments (i.e., the arrangement of words) and the coherence between said argument and the topic it relates to. The authors acknowledge that the two aspects may be representative of the cogency, reasonableness and clarity, and demonstrate that "incorporating both syntactic and coherence information can boost the classification performance compared to the models without considering them". Thus, in recent years the focus seems to be moving towards an even more targeted fine-grained analysis of the dimensions, allowing not only to rate a single quality aspect (e.g., appropriateness, sentiment), but to consider potentially related factors (e.g., toxic emotions, aggressiveness) (Ziegenbein et al., 2023; Falk et al., 2024). With the development of a commonly used coding scheme, Wachsmuth et al. (2017a) address one "key requirement in enhancing reusability" (Reed et al., 2008). The authors state that they do "not propose a specific approach to assess quality; rather [define] a common ground by providing a (...) holistic view" (Wachsmuth et al., 2017a).

Argument Quality Taxonomy. Due to the ongoing widespread use of Wachsmuth et al. (2017a)'s taxonomy in the domain, we base our overview of quality dimensions on their proposed taxonomy and extend it with further dimensions identified in the surveyed papers[3]. Figure 3 depicts all of the quality dimensions found in the surveyed papers. Here, it is essential to point out that it is far beyond the scope of this survey to propose a novel or updated taxonomy. Yet, we believe that developments in the field, similarly to such in the field of philosophy, would sooner or later inevitably lead to adaptations to the first taxonomy proposed by Wachsmuth et al. (2017a).

The taxonomy differentiates between three high-level quality dimensions. *Logic / Cogency* assesses whether the premises within an argument are acceptable, relevant and sufficient to its conclusion. *Dialectic / Reasonableness* explores whether an argumentation would be accepted by the target audience, whether it is relevant to the issue and sufficient against counter-arguments. *Rhetoric / Effectiveness* addresses the persuasiveness of an argumentation towards a target audience in terms of the credibility of the author, the emotional appeal, the style and the appropriateness of the used language, and the arrangement of the argumentation. The high-level quality dimensions are then split into sub-dimensions, which look at individual quality aspects more precisely. The interested reader may refer to the work of Wachsmuth et al. (2017a) for the exact definitions of the individual (sub-)dimensions.

Throughout the creation of this taxonomy, Wachsmuth et al. (2017a) analyse existing approaches to AQ, yet many of them are not included within the final taxonomy. Instead the authors leave out some dimensions such as persuasiveness as they are deemed too close to a high-level dimension and add references to further quality dimensions previously addressed by the AQ literature. Our overview includes the taxonomy by Wachsmuth et al. (2017a), all quality dimensions explored throughout the creation of the taxonomy, as well as further dimensions discovered as part of this survey (i.e., sentiment, objectivity, impact, strength). As such we aim to recognize their frequent appearance and relevance in the domain (e.g., Lukin et al., 2017; Shiota and Shimada, 2020).

Related Fields. In our analysis we categorize work targeting the identification of argument components as AM and work targeting the comparison between versions of the same argumentation as AImp. Some prior work has built connections between the fields. The work of Li et al. (2020) and Liu et al. (2021) for instance bridges the focus points of AM and AQ by examining the relation between the discourse structure of arguments to persuasiveness. Others describe the analysis of the structure of the arguments as "the first step in analysing [argument] quality." Our overview acknowledges the close connection and oftentimes blurred boundaries between AQ, AM and AImp, while taking a closer look at the individual dimensions that are relevant to the quality of the argumentation itself. Further, it may be argued that Persing et al.'s work (2010; 2013; 2014; 2015) on student essays is more related to Automated Essay Scoring (AES) than AQ, yet it has (also) been widely accepted by prior work as a contribution to the assessment of argumentation quality. The boundary between domains is further blurred when essays are evaluated based on their argumentative structure and AM techniques are applied (Stab and Gurevych, 2014).

The interested reader may refer to Lawrence and Reed (2020) and Vecchi et al. (2021) for a survey on AM and to Ramesh and Sanampudi (2022) for a survey on AES. For the first survey on AQ providing a "holistic view on argumentation quality assessment in natural language" see Wachsmuth et al. (2017a).

4 Annotated Datasets

We found a total of 57 datasets which either extended existing datasets with further annotations or were created from scratch. The main purposes that the datasets target are *Argument Mining* (Stab and Gurevych, 2014; Habernal and Gurevych, 2017; Shnarch et al., 2018), *Argumentation Quality*, *Argument Improvement* (Zhang et al., 2017; Afrin and Litman, 2018; Skitalinskaya et al., 2021), and *Argument Search* (Wachsmuth et al., 2017b; Stab et al., 2018; Nilles et al., 2021). As our focus lays on the field of AQ, we take a closer look at the 32 directly relevant datasets of the total surveyed datasets in regards to their size, language, quality dimensions, annotation scales, annotation process, and identified annotation issues. The selection of the criteria was based on previous studies on annotated datasets from related domains (e.g., Van

[3] Note that we exclude sub-dimensions which are repetitive in their naming such as "thesis clarity" as a sub-category of "clarity". This differentiation is not essential in our case as individual related publications are discussed in Section 4.

Der Lee et al., 2019; Ke and Ng, 2019).

Our overview is based on the datasets, their annotation guidelines and the publications introducing them. The majority of publications uploaded supplemental material or made it easy to discover or request the datasets. In 2 out of the 7 attempts to reach the authors of the respective papers in order to request the datasets remained without a response, and for 1 publication none of the authors' emails are active anymore.

4.1 Size and Language

The authors of the individual datasets have taken various approaches to the evaluation of argumentation quality. As such they have also used various text entities to measure the size of the final annotated datasets[4] (e.g., 374 472 comments, 320 arguments, 830 essays). The *text entities* include: the number of propositions, premises, claims, arguments, words, sentences, essays, (online) posts, (discussion/forum) threads, and evidence/argument pairs. The wide variability in the annotated entity types and thus the granularity of the annotated quality dimensions makes a direct comparison of the dataset sizes quite difficult. A detailed overview of all datasets is depicted in Table 1 in Appendix A.

In regards to the languages of the datasets, 3 (9.38%) of the publications explicitly state the language, 28 (87.5%) implicitly indicate it (e.g., through the provided examples) or we were able to detect the language by taking a look at the respective published or received-on-request datasets. We could not deduct the language of the datasets described in one publication. One dataset, which was created for AM, yet includes an overall AQ score, is multi-lingual (Toledo-Ronen et al., 2020). All of the remaining datasets were created from argumentation in English.

Research Gaps and Future Work. While English is generally over-represented in the NLP community in terms of datasets, models, tools, etc., we believe that the numbers here present a far outlier. One of the limitations of our approach is the initial use of the English term for the search for publications. However, it is surprising that even within the 211 discovered publications, we did not find any dataset that is created in another language (with the potential exception of the one unknown dataset). We would like to point out that this is not necessarily the case for AM, where few datasets exist also for other languages (e.g., Wambsganss and Niklaus, 2022). Future work should address this research gap by exploring non-English argumentative texts.

Opposing statements were made by prior research in regards to the length of text entities in the annotation process. Swanson et al. (2015) and Gienapp et al. (2020) recognize a negative correlation of annotation quality and length of text entity. In contrast, Wachsmuth and Werner (2020) observe better judgment for longer arguments. Joshi et al. (2023) take a closer look at the annotation scores for various text lengths and discover a normally distributed curve with "a peak score from 210-270 characters", contributing it to the idea that too few characters may be insufficient to make a persuasive point, yet too many may also be considered not persuasive. Yet, future work should explore whether this ideal length is dependent on the text type (e.g., news article vs. online forum) or on the annotated level of granularity (i.e., premise/conclusion, argument, or argumentation).

4.2 Quality Dimensions

In the majority of the publications describing the creation of datasets, the annotated quality dimensions are either explicitly stated or can be deducted from the publication or supplemental material. A detailed overview of the surveyed datasets for AQ with their respective annotated quality dimensions can be found in Appendix A. Here, we only included the argument aspects annotated for AQ and excluded those better fitting to related areas (e.g., refutation method in Wei et al., 2016). The dimensions used for the creation of the overview are a subset of the ones described in Figure 3.

In some of the cases, however, the assignment of the described quality aspects to a specific quality dimension may be ambiguous. El Baff et al. (2018) explore "whether an editorial brings readers of opposing belief closer together or rather increases the gap between them". Here, the quality perspective is not clearly covered by a single sub-dimension. Therefore we resort to the general categorization made by the authors themselves and assign the dataset to the high-level dimension *reasonableness* (i.e., dialectic). The dataset created by Tan et al. (2016) is another such example. The authors collect Reddit[5] data where an original poster (OP) asks other users to change their view on a topic and make use of a parameter indicating whether a par-

[4] Note that the final size oftentimes differs from the initial raw dataset size.

[5] https://reddit.com/r/changemyview

ticular response changed their opinion (i.e., delta label) to explore *persuasion* in online discussions. However, due to the contrasting opinions of the OP and the responding user, the label can be viewed as an indication of *global acceptability*.

Toledo et al. (2019), Gretz et al. (2020), and Joshi et al. (2023) explore whether the annotators would "recommend" to use an argumentation in a speech. We believe that this aspect is most closely connected to the *global relevance* of an argument (i.e., "it contributes to the issue's resolution" (Wachsmuth et al., 2017a)). On the contrary, we categorize the measure of a "relevance level" in Dumani and Schenkel (2019) as *local relevance*, as it explores the relevance of claims in pairs of ⟨query claim, result claim⟩.

Zhang et al. (2016b) take a look at the different styles of argumentation in moderated live debates. For each debate the "winning side" is described as the higher delta between the received votes pre and post-debate. Here our approach aligns with Wachsmuth et al. (2017a) as we do not further categorize this dataset into a sub-dimension.

The majority of the surveyed datasets are based on text from essays, online debate portals, forums, news articles, but only few of them originate from a more specific domain. Alhamzeh (2023) created a dataset consisting of 80 quarterly organized events of public traded companies. The annotated dimensions included generally relevant quality dimensions for argumentation (e.g., strength, specificity, persuasiveness), but also specific ones targeting the financial domain (e.g., Which quarter/year does the argument refer to?). One of the annotated aspects states whether an argument is "objective" or not, which seems closely related to *emotional appeal* (Wachsmuth et al., 2017a).

Research Gaps and Future Work. The surveyed literature (Wachsmuth et al., 2017a; Habernal and Gurevych, 2016b; Liu et al., 2023) oftentimes refers to existing argumentation theories such as Toulmin (1958)'s model for the quality of the general argument structure. Prior work has taken a closer look at the relation between the use of Toulmin's model and the quality of an argument. To the authors' surprise "[not] only did the Toulmin model create arguments with decreased clarity, but it also decreased the personal relevance, sense of urgency, and drastically decreased the overall level of agreeability" (Dorton et al., 2021). This raises questions as to whether the currently used quality dimensions are the best suit for the quality assessment. Future work should take a closer look at the suitability of current approaches for the task at hand and explore further related perspectives on argumentation quality.

4.3 Absolute vs. Relative Quality

Prior work mostly describes individual approaches and datasets based on the quality dimensions that they analyze. In addition, we consider whether the individual quality dimensions were regarded in a relative or in an absolute manner following the distinction made by Toledo et al. (2019).

The *relative* quality analysis evaluates the relation between pairs of text entities instead of regarding them as individual statements. The most frequently observed type of relative evaluation is the preference comparison of an argument A over an argument B (Habernal and Gurevych, 2016b; Gleize et al., 2019; Toledo et al., 2019; Gienapp et al., 2020). This approach is used to reduce the annotation complexity by requiring no prior knowledge from the annotators (Gienapp et al., 2020). While the relative quality offers less specific evaluation, it allows for a *new best argument* to be defined at all time (i.e., as it is simply tagged as better than the previous best) (Dumani and Schenkel, 2020). Another common evaluation is done by categorizing the relation between pairs of text entities (e.g., Wei et al., 2016; Habernal and Gurevych, 2016a).

From an *absolute* point of view arguments are analyzed as individual text entities (e.g., single argument) or in conjunction with further related text entities (e.g., a topic and an argument). Individual text entities such as sentences are then analyzed in terms of their organization, sentiment, clarity, strength, relevance, sufficiency, persuasiveness, winning side, reasonableness, or a mixture of multiple quality dimensions (see Appendix A for a complete overview). Overall 23 publications took an absolute approach, 6 took a relative one, and Toledo et al. (2019) applied both in their work, creating two distinct datasets. This categorization (i.e., absolute vs. relative) is not to be confused with the differentiation between intrinsic and extrinsic quality dimensions. An intrinsic evaluation is based only on the text of the argument (e.g., Wachsmuth and Werner, 2020), while an extrinsic one requires previous knowledge such as background or context (e.g., Potash et al., 2017).

Research Gaps and Future Work. Gienapp et al. (2020) distinguish between rating methods that use an interval scale (e.g., Likert scale) and "relative comparison", where annotators view two texts at a time and are asked to state their preference (i.e., relative quality) in terms of the argumentation quality. They observe better overall inter-annotator agreements when a relative comparison is used. This is assumed to be related to two major drawbacks of absolute rating scales in this context. On the one hand, the use of an interval scale may lead to incorrect conclusions based on statistical methods. This is because "assessors rarely perceive labels as equidistant, thus producing only ordinal data [which] leads to a misuse of statistical tests and results in low statistical power of subsequent analyses" (Gienapp et al., 2020). On the other hand, such rating has proven to be difficult for annotators without previous knowledge. This claim is supported by the results of the annotation by Lauscher et al. (2018), where the various inter-annotator agreement scores "suggest that the difficulty of the task is highly dependent on the domain".

In general, higher quality datasets are required in order for tools to be able to perform better. We acknowledge that relative comparison yields better agreements, yet also recognize that over 75% of all datasets are annotated in an absolute manner, which may be an indication of its better suitability for further use. Therefore, we suggest that future work explores ways to translate relative annotations to absolute ones. In addition, one could address the absolute quality assessment in particular with an aim to better understand the difficulties in annotation from the point of view of the annotators.

4.4 Annotation Scales

Across the surveyed datasets we find annotation scales of various types. While *relative* quality is mostly measured by stating a preference of one argument over another (Toledo et al., 2019; Gleize et al., 2019; Gienapp et al., 2020), Habernal and Gurevych (2016b) included an option where both arguments are equally convincing.

Previous attempts have been made to refine the coarse granularity of the relative annotations. Habernal and Gurevych (2016b) apply PageRank on a directed acyclic graph derived from their annotated data. Chen et al. (2013) introduce an online sampling method based on the Bradley-Terry model (Bradley and Terry, 1952). The shortcomings of online sampling methods for crowdsourcing (i.e., not allowing multiple workers to annotate simultaneously or to not have a preference) have been addressed by Gienapp et al. (2020) in an offline sampling method, which produces scalar ranking scores from the preference annotations. Britner et al. (2023) point out that no attention has been given to justifying why a certain argument is predicted to be better than another and introduce an application which addresses this gap. Their approach makes use of various absolute quality dimensions.

Annotation scales following the *absolute* approach have a higher variety. A point scale is used in 48% of the surveyed datasets, in which arguments are evaluated individually. However, even within these datasets different ranges for scales are used: 1-3, 0-2, 1-5, 1-6, 1-4 at half point increments, -5 to 5. Also here, there are few cases (e.g., Dumani and Schenkel, 2020) where the annotators are allowed to state that they "cannot judge". Marro et al. (2022) initially chose an interval scale (i.e., 0, 5, 10, 15, 20, 25), yet switched to an ordinal scale (i.e., 0, 15, 25) to achieve higher annotation quality.

In addition to an interval scale, the annotators in Persing and Ng (2017a,b) were asked to identify "five errors that could have a negative impact on (...) persuasiveness", while Alhamzeh (2023) included further domain-specific dimensions with a categorical scale. Similarly, Falk et al. (2024) assessed sentiment via three categories. Durmus et al. (2019) took a different approach by using labels: no impact, low impact, medium impact, high impact, very high impact, which in contrast to the aforementioned examples are not assigned to interval-scale values.

Some annotations (e.g., Swanson et al., 2015; Falk et al., 2024) took a simplified approach and used binary alternatives for the tagging of arguments. Similarly to the refining of the coarse-grained relative annotation approaches, Toledo et al. (2019), Gretz et al. (2020) and Toledo-Ronen et al. (2020) convert the binary tags to a more precise quantitative score value between 0 and 1 after the annotation has been completed. Swanson et al. (2015) skip this step by using a slider from 0 to 1, simplifying the task for the annotators, yet preserving the finer granularity of the evaluations.

Research Gaps and Future Work. The analysis of the rating scales used for the annotation of the surveyed datasets shows that point scales are most frequently used. However, the variety of their range and the different annotation guidelines often lead to different meanings behind the same numbers. Due to the high costs associated with the annotation of datasets, future work should explore options to port the different scales in an aim to increase the reusability of already existing datasets.

4.5 Annotation Process

One of the most frequently chosen annotation approaches is crowdsourcing (e.g., Amazon's Mechanical Turk, Figure-Eight) (Wachsmuth et al., 2014; Swanson et al., 2015; Habernal and Gurevych, 2016a,b; Braunstain et al., 2016; Shnarch et al., 2018; Gleize et al., 2019; Toledo et al., 2019; Ng et al., 2020; Gretz et al., 2020; Gienapp et al., 2020). Another choice is the use of a graphical user interface (El Baff et al., 2018; Dumani and Schenkel, 2019, 2020; Alhamzeh, 2023). Due to the vastly different approaches for the annotation task, the number of annotators also varies accordingly. With crowdsourcing the number of annotators per entity varied from 3 to 17 and per dataset from 90 to 3 900. Oftentimes, certain criteria were applied to ensure the quality of the annotators' work (e.g., native or proficient speakers, having a high acceptance rate of their previous annotations). When other approaches were chosen, the number of annotators varied between 2 and 8. In these cases the criteria for the annotators selection could be made more precisely. While the language proficiency is also a factor here, the criteria further included among others a linguistic background, an expertise in AM, or an expertise in the arguments' domain.

The annotation quality is predominantly evaluated via inter-annotator agreement scores. The majority of the datasets calculated Cohen's kappa between pairs of annotators (0.322 - 0.848), Fleiss' kappa (0.457 - 0.86) or Krippendorff's alpha (0.00 - 0.935). A few authors (e.g., Toledo et al., 2019; Gretz et al., 2020; Falk et al., 2024) included further techniques with the aim to increase the quality of the annotations such as adding test questions, organizing small pilot annotations, offering annotators a test run to familiarize them with the task.

Various quality levels of annotators' work may be addressed in the post-processing of the data by applying scoring functions for annotations. MACE probability (Hovy et al., 2013) uses a generative model to estimate the true label and annotator reliability (Habernal and Gurevych, 2016b,a; Joshi et al., 2023), while Weighted Average factors in the annotator reliability weight their judgments as means to reduce the influence of non-reliable annotators on the final quality score (Gretz et al., 2020; Joshi et al., 2023). Simpler methods include the use of a majority agreement, full agreement or similar (Persing et al., 2010; Persing and Ng, 2013, 2015, 2017b,a; El Baff et al., 2018; Wachsmuth et al., 2017a).

Research Gaps and Future Work. Overall, the achieved inter-annotator scores oftentimes indicated that the tasks are difficult for humans (Persing et al., 2010; Gleize et al., 2019; Ng et al., 2020). Dumani and Schenkel (2020) identify the nominal scores for reasonableness as having the highest level of disagreement. Stab and Gurevych (2017) link the use of modal verbs (e.g., can) and unspecific quantifiers (e.g., some, many, various) to a decrease in the agreement among annotators, suggesting to address the issues by providing more precise annotation guidelines. Alhamzeh (2023) connects the later issue to a perception of "low degrees of specificity, strength, and persuasiveness".

Further, subjectivity is also recognized as a reason for low inter-annotator agreement. Habernal and Gurevych (2016a) suggest that some quality dimensions may require a description of the target audience due to their subjective nature. Wei et al. (2016) found some sub-categories in their annotation to be difficult to distinguish (e.g., target losing argument and refutation), leading to mismatches in the annotation. Ng et al. (2020) observe a higher disagreement in cases where a particular topic is "deemed 'less worthy' of being discussed, and (...) humorous in nature or had trivial consequences." In addition, statements including sarcasm, irony or rhetorical questions are deemed difficult to annotate. Future work should look into better suited annotation approaches for these particular issues in the annotation process.

5 Further Aspects and Related Fields

Falk and Lapesa (2023) differentiate between the AQ perspectives discussed in the AM community and those in the Deliberative Theory. The former focus "on the logical dimension or specific aspects of persuasion", while the later "puts the discourse as a whole and the interaction between discourse participants into the focus". Future work should explore the suitability, the overlap, and the compatibility of the two perspectives, as the progress made in the one domain could potentially contribute to the work in the other.

The recent survey by Guerraoui et al. (2023) discusses feedback systems specifically for argumentation by categorizing argument feedback into four categories - Richness, Visualization, Interactivity, Personalization. They note that future research should focus on considering the author's skill level for the feedback. We follow this statement and suggest to consider the personal background (e.g., school level, native language) of an argument's author also when measuring the argumentation quality. In the surveyed datasets the education level, setting and personal skill level of the authors of the argumentative texts were typically not stated. Reed et al. (2008) point out that restricting the goal of a corpus "is to permanently restrict the scope of what they can support". Further, Kasneci et al. (2023) note that training data for Large Language Models (LLM) should be diverse to reduce bias towards any particular group. We believe that this is also essential for the training of any models in the AQ domain to ensure a more versatile and precise assessment. When we consider the various philosophical and cultural view points and understandings of what a good argument is (e.g., Perelman, 1971, Wenzel, 1990), such richness in the datasets could reduce the risk of creating a very narrow view on AQ within the NLP community.

Gienapp et al. (2020) describes the varying reference frames of crowdsourcing annotators as an issue due to its negative effect on the inter-annotator agreement scores. On the contrary, Plank (2022) points out that "a crucial assumption of today's learning systems is to rely on a single gold label per instance", which disregards the various opinions and subjective interpretation of annotators when language is involved. This issue is transferable to the assessment of AQ as the task is also subjective. While the issue of ambiguity in gold standards (Poesio and Artstein, 2005) is not new, it is worth considering whether an alternative annotation format such as multi-layer labeling (e.g., Bamman et al., 2019 in Named Entity Recognition for English literature) could also be beneficial to the AQ domain. This aspect is essential to argumentation, as the perceived quality of an argument is oftentimes influenced by external factors such as the personal view of an annotator or their familiarity with a topic, to name a few.

The evaluation of AQ can be explored beyond the textual form. Previous work has looked at other types of media such as video and audio to detect trembling in the voice, gesticulation, face expressions of participants during debates (e.g., Shiota and Shimada, 2020; Hasan et al., 2021). While this research field is one of the most closely related to the quality of textual argument, it is beyond the scope of this paper to discuss the various aspects of behavioral analysis and its potential correlation to argumentation quality. Yet, considering the fact that text is merely one of the attributes of in-person discussions, it is worth exploring whether the tools and datasets created for the analysis of written text can also be applied to texts extracted from face-to-face discussions.

6 Conclusion

While the AQ domain is multi-faceted, posing a complex challenge, prior research has tackled it from various perspectives. The interdisciplinary nature of the quality assessment task offers the opportunity to gain knowledge from related research fields, while also being able to contribute back to the related fields e.g., through the creation of annotated datasets and the evaluation of hypotheses about the significance or relations of the individual quality dimensions. We present a survey of existing argumentation quality perspectives for computational assessment and annotated datasets created or suitable for the domain. We outline potential shortcomings and research gaps from prior work, and suggest future work that may be beneficial to the further development of approaches and tools.

7 Limitations

The initial search for related work used a keyword which is quite general, yet we reduced the scope to publications in the field of computer science by limiting the initial collection to the DBLP library. To compensate for this shortcoming, we applied the Snowballing approach, which allowed us to increase the number of discovered publications almost 3 times. Nevertheless, our approach does not guarantee that less frequently cited publications which may be related to the topic have not remained undiscovered.

Further, the overview of the annotated datasets was created based on the surveyed work. In our case all but one available and received (on request) datasets annotated arguments in English. Therefore, despite our aim to cover the domain of Argumentation Quality as representatively as possible, we cannot guarantee that there are no other branches of the domain that target other languages.

References

Tazin Afrin and Diane Litman. 2018. Annotation and classification of sentence-level revision improvement. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 240–246.

Alaa Alhamzeh. 2023. Financial argument quality assessment in earnings conference calls. In *International Conference on Database and Expert Systems Applications*, pages 65–81. Springer.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. 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)*, pages 2138–2144.

J Anthony Blair. 2011. *Groundwork in the theory of argumentation: Selected papers of J. Anthony Blair*, volume 21. Springer Science & Business Media.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Liora Braunstain, Oren Kurland, David Carmel, Idan Szpektor, and Anna Shtok. 2016. Supporting human answers for advice-seeking questions in cqa sites. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*, pages 129–141. Springer.

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*, pages 5015–5020.

Xi Chen, Paul N Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 193–202.

Stephen L Dorton, Samantha B Harper, Glory A Creed, and H George Banta. 2021. Up for debate: Effects of formal structure on argumentation quality in a crowdsourcing platform. In *International Conference on Human-Computer Interaction*, pages 36–53. Springer.

Lorik Dumani and Ralf Schenkel. 2019. A systematic comparison of methods for finding good premises for claims. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 957–960.

Lorik Dumani and Ralf Schenkel. 2020. Quality-aware ranking of arguments. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 335–344.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019. The role of pragmatic and discourse context in determining argument impact. 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 5668–5678.

Frans H. van Eemeren, Rob Grootendorst, Ralph H. Johnson, Christian Plantin, Charles A. Willard, Rob Grootendorst, Ralph H. Johnson, Christian Plantin, and Charles A. Willard. 1996. *Fundamentals of Argumentation Theory*. Routledge.

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.

Neele Falk and Gabriella Lapesa. 2023. Bridging argument quality and deliberative quality annotations with adapters. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2424–2443.

Neele Falk, Eva Maria Vecchi, Iman Jundi, and Gabriella Lapesa. 2024. Moderation in the wild: Investigating user-driven moderation in online discussions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 992–1013.

Andrea Galassi, Marco Lippi, and Paolo Torroni. 2023. Multi-task attentive residual networks for argument mining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1877–1892.

Lukas Gienapp, Benno Stein, Matthias Hagen, and Martin Potthast. 2020. Efficient pairwise annotation of argument quality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5772–5781.

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.

Leo A Goodman. 1961. Snowball sampling. *The annals of mathematical statistics*, pages 148–170.

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 AAAI Conference on Artificial Intelligence*, volume 34, pages 7805–7813.

Camelia Guerraoui, Paul Reisert, Naoya Inoue, Farjana Sultana Mim, Keshav Singh, Jungmin Choi, Irfan Robbani, Shoichi Naito, Wenzhi Wang, and Kentaro Inui. 2023. Teach me how to argue: A survey on NLP feedback systems in argumentation. In *Proceedings of the 10th Workshop on Argument Mining*, pages 19–34, Singapore. Association for Computational Linguistics.

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.

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.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Md Kamrul Hasan, James Spann, Masum Hasan, Md Saiful Islam, Kurtis Haut, Rada Mihalcea, and Ehsan Hoque. 2021. Hitting your marq: Multimodal argument quality assessment in long debate video. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6387–6397.

Diana E Hess. 2009. *Controversy in the classroom: The democratic power of discussion*. Routledge.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Ioana Hulpus, Jonathan Kobbe, Christian Meilicke, Heiner Stuckenschmidt, Maria Becker, Juri Opitz, Vivi Nastase, and Anette Frank. 2019. Towards explaining natural language arguments with background knowledge. In *PROFILES/SEMEX@ ISWC*, pages 62–77.

Omkar Joshi, Priya Pitre, and Yashodhara Haribhakta. 2023. Arganalysis35k: A large-scale dataset for argument quality analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13916–13931.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

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*. International Joint Conferences on Artificial Intelligence Organization.

Anne Lauscher, Lily Ng, Courtney Napoles, Maria Nadejde, Joel Tetreault, Junchao Zheng, Courtney Napoles, Joel Tetreault, Courtney Napoles, Chris Callison-Burch, et al. 2018. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics*, volume 7, pages 229–234. International Committee on Computational Linguistics.

Anne Lauscher, Henning Wachsmuth, Iryna Gurevych, and Goran Glavaš. 2022. Scientia potentia est—on the role of knowledge in computational argumentation. *Transactions of the Association for Computational Linguistics*, 10:1392–1422.

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

Jialu Li, Esin Durmus, and Claire Cardie. 2020. Exploring the role of argument structure in online debate persuasion. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8905–8912.

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2021. Exploring discourse structures for argument impact classification. 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)*, pages 3958–3969.

Zhexiong Liu, Diane Litman, Elaine Wang, Lindsay Matsumura, and Richard Correnti. 2023. Predicting the quality of revisions in argumentative writing. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 275–287, Toronto, Canada. Association for Computational Linguistics.

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.

Santiago Marro, Elena Cabrio, and Serena Villata. 2022. Graph embeddings for argumentation quality assessment. In *EMNLP 2022-Conference on Empirical Methods in Natural Language Processing*.

Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, and Kohsuke Yanai. 2022. End-to-end Argument Mining with Cross-corpora Multi-task Learning. *Transactions of the Association for Computational Linguistics*, 10:639–658.

Lily Ng, Anne Lauscher, Joel Tetreault, and Courtney Napoles. 2020. Creating a domain-diverse corpus for theory-based argument quality assessment. In *Proceedings of the 7th Workshop on Argument Mining*, pages 117–126.

Markus Nilles, Lorik Dumani, and Ralf Schenkel. 2021. Quark: A gui for quality-aware ranking of arguments. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2546–2549.

Chaim Perelman. 1971. *The New Rhetoric*. Springer.

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.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269.

Isaac Persing and Vincent Ng. 2014. Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1534–1543.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552.

Isaac Persing and Vincent Ng. 2017a. Lightly-supervised modeling of argument persuasiveness. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 594–604.

Isaac Persing and Vincent Ng. 2017b. Why can't you convince me? modeling weaknesses in unpersuasive arguments. In *IJCAI*, pages 4082–4088.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682.

Moritz Plenz, Raphael Buchmüller, and Alexander Bondarenko. 2023a. Argument quality prediction for ranking documents. *Working Notes of CLEF*.

Moritz Plenz, Juri Opitz, Philipp Heinisch, Philipp Cimiano, and Anette Frank. 2023b. Similarity-weighted construction of contextualized commonsense knowledge graphs for knowledge-intense argumentation tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6130–6158, Toronto, Canada. Association for Computational Linguistics.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 76–83, Ann Arbor, Michigan. 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.

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie-Francine Moens. 2008. Language resources for studying argument. In *Proceedings of the 6th conference on language resources and evaluation-LREC 2008*, pages 2613–2618. ELRA.

Tsukasa Shiota and Kazutaka Shimada. 2020. The discussion corpus toward argumentation quality assessment in multi-party conversation. In *2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 280–283. IEEE.

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2018. Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605.

Gabriella Skitalinskaya, Jonas Klaff, and Henning Wachsmuth. 2021. Learning from revisions: Quality assessment of claims in argumentation at scale. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1718–1729.

Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. Argumentext: Searching for arguments in heterogeneous sources. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: demonstrations*, pages 21–25.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 1501–1510.

Christian Stab and Iryna Gurevych. 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.

Xichen Sun, Wenhan Chao, and Zhunchen Luo. 2021. Syntax and coherence-the effect on automatic argument quality assessment. In *Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part II 10*, pages 3–12. Springer.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue*, pages 217–226.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624.

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.

Orith Toledo-Ronen, Matan Orbach, Yonatan Bilu, Artem Spector, and Noam Slonim. 2020. Multilingual argument mining: Datasets and analysis. In *Findings of the association for computational linguistics: Emnlp 2020*, pages 303–317.

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

Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2020. Fine-grained argument unit recognition and classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9048–9056.

Chris Van Der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368.

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. 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)*, pages 1338–1352.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017a. 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.

Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017b. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014. A review corpus for argumentation analysis. In *Computational Linguistics and Intelligent Text Processing: 15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6-12, 2014, Proceedings, Part II 15*, pages 115–127. Springer.

Henning Wachsmuth and Till Werner. 2020. Intrinsic quality assessment of arguments. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6739–6745.

Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *LREC*, volume 12, pages 812–817. Istanbul, Turkey.

Thiemo Wambsganss and Christina Niklaus. 2022. Modeling persuasive discourse to adaptively support students' argumentative writing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8748–8760.

Zhongyu Wei, Yandi Xia, Chen Li, Yang Liu, Zachary Stallbohm, Yi Li, and Yang Jin. 2016. A preliminary study of disputation behavior in online debating forum. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 166–171.

Joseph W Wenzel. 1990. Three perspectives on argument: Rhetoric, dialectic, logic. *Perspectives on argumentation: Essays in honor of Wayne Brockriede*, pages 9–26.

Fan Zhang, Homa B Hashemi, Rebecca Hwa, and Diane Litman. 2017. A corpus of annotated revisions for studying argumentative writing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1568–1578.

Fan Zhang, Rebecca Hwa, Diane Litman, and Homa B Hashemi. 2016a. Argrewrite: A web-based revision assistant for argumentative writings. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Demonstrations*, pages 37–41.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016b. Conversational flow in oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141.

Timon Ziegenbein, Shahbaz Syed, Felix Lange, Martin Potthast, and Henning Wachsmuth. 2023. Modeling appropriate language in argumentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4344–4363.

A Annotated Datasets for Computational Argumentation Quality Assessment

The following section gives a tabular overview of the annotated datasets, which we discovered through this survey on the domain of Computational Argumentation Quality Assessment. In some cases, the datasets have been given a *name*. When this is not the case, they can be clearly identified based on their authors' *reference* and publication *year*. Next, the *size* and the annotation *approach* are presented. While some *quality dimensions* are mentioned explicitly in the respective descriptions of the datasets or can be easily deducted, for others we categorized them in one of the quality dimensions previously described in the literature. In these cases, the assumed category is added in brackets and in an italic font next to the originally described quality dimension. Following, we list the *rating scales* used for the annotation of the respective dimensions. Lastly, the *availability* is given as "Yes", "On Request", or "Not Reachable". The latter is a case where the dataset is not shared, cannot be found online, and none of the authors' emails are active anymore.

[6] Name given by Gretz et al. (2020)

[7] The dataset was described and used for AQ by Zhang et al. (2016b). The dataset was created as part of public debates organized by Intelligence Squared.

Table 1: Overview of annotated datasets for the domain of computational argumentation quality assessment including their name (if given), authors, publication year, size, approach (i.e., absolute or relative), annotated quality dimension(s), annotation rating scale(s), and availability.

Name	Reference	Year	Size	Approach	Quality Dimension	Rating Scale	Availability
	Persing et al.	2010	1 003 essays	absolute	organization	1-4 at half-point	Yes
Internet Argument Corpus	Walker et al.	2012	10 003 pairs 6 797 chains of three posts	absolute	emotional appeal (among others)	from -5 to 5 or "can't tell"	Yes
	Persing and Ng	2013	830 essays	absolute	thesis clarity	1-4 at half-point	Yes
	Persing and Ng	2014	830 essays	absolute	prompt adherence	1-4 at half-point	Yes
ArguAna TripAdvisor Corpus	Wachsmuth et al.	2014	31 006 statements 24 596 product features	absolute	sentiment	fact, positive opinion, negative opinion	Yes
	Persing and Ng	2015	1 000 essays	absolute	strength	1-4 at half-point	Yes
SwanRank[6]	Swanson et al.	2015	109 074 posts	absolute	overall	slider from 0 to 1	Yes
	Braunstain et al.	2016	5 000 sentences	absolute	emph(local) relevance, level of support	ordinal, binary (derived from an ordinal scale)	Yes
CMV	Tan et al.	2016	20 626 discussion threads	absolute	delta label *(persuasiveness)* (among others)	binary	Yes
	Wei et al.	2016	45 argument pairs	relative	overall (among others)	categories	Not Reachable
UKPConvArg1	Habernal and Gurevych	2016b	11 650 arguments	relative	convincingness	A>B, B>A, A=B	Yes
UKPConvArg2	Habernal and Gurevych	2016a	9 111 argument pairs	relative	convincingness	decision-tree based classification	Yes
Intelligence Squared Debates	Zhang et al.[7]	2016b	108 debates	absolute	winning side	percentage of votes	Yes
	Persing and Ng Persing and Ng	2017b 2017a	1 208 comments	absolute	persuasiveness	1-6; categories for error types	Yes
UKP Insufficient Arguments	Stab and Gurevych	2017	1 029 arguments	absolute	*(local)* sufficiency	binary	Yes

Name	Reference	Year	Size	Approach	Quality Dimension	Rating Scale	Availability

Dagstuhl-15512 ArgQuality Corpus	Wachsmuth et al.	2017a	320 arguments	absolute	cogency (l. acceptability, l. relevance, l. sufficiency);
					effectiveness (credibility, emotional appeal, clarity, appropriateness, arrangement);
					reasonableness (g. acceptability, g. relevance, g. sufficiency);
					overall
						1-3 or “cannot judge”	Yes

Webis-Editorial-Quality-18 Corpus	El Baff et al.	2018	1 000 news editorials	absolute	reasonableness	categories	Yes
	Durmus et al.	2019	47 219 claims	absolute	impact	categories	On Request
IBM-EviConv	Gleize et al.	2019	5 697 evidence pairs	relative	convincingness	A>B, B>A	Yes
IBM-Rank	Toledo et al.	2019	5 298 arguments	absolute	recommend, convincingness	binary	Yes
IBM-Pairs	Toledo et al.	2019	9 100 pairs	relative	recommend, convincingness	A>B, B>A	Yes
	Dumani and Schenkel	2019	7 444 pairs	relative	*(local)* relevance	categories	On Request
	Dumani and Schenkel	2020	1 376 premises	absolute	cogency, reasonableness, effectiveness	1-3 or “cannot judge”	On Request
Webis-ArgQuality20	Gienapp et al.	2020	41 859 pairs	relative	logic, dialectic, rhetoric	A>B, B>A	Yes
IBM-ArgQ-Rank-30kArgs	Gretz et al.	2020	30 497 arguments	absolute	recommend *(global relevance)*	binary	Yes
	Toledo-Ronen et al.	2020	30 497 arguments	absolute	overall	binary	Yes

Name	Reference	Year	Size	Approach	Quality Dimension	Rating Scale	Availability
GAQ Corpus	Ng et al.	2020	5 285 arguments	absolute	logic, dialectic, rhetoric, overall	1-5 or “cannot judge”	On Request
	Marro et al.	2022	1 908 arguments	absolute	cogency, reasonableness, rhetoric	ordinal scale	On Request

FinArgQuality	Alhamzeh	2023	14 146 sentences	absolute	strength,
					specificity,
					persuasiveness,
					objective *(emotional appeal)*,
					temporal-history
						0-2;
						binary;
						categorical
							Yes

ArgAalysis35K	Joshi et al.	2023	35 000 argument-analysis pairs	absolute	recommend *(global relevance)*	binary	On Request

Appropriateness Corpus	Ziegenbein et al.	2023	2 191 arguments	absolute	(in)*appropriateness*;
					toxic emotions (excessive intensity, emotional deception);
					missing commitment (missing seriousness, missing openness);
					missing intelligibility (unclear meaning, missing relevance, confusing reasoning);
					other reasons (detrimental orthography, reason unclassified)
						ordinal;
						binary;
							Yes

UMOD	Falk et al.	2024	1 000 comment-reply pairs	absolute	subjectivity,
					agressiveness,
					constructiveness,
					sentiment
					(among others)
						1-5;
						binary;
						categorical
							Yes

B References to Surveyed Corpus

In this list of references we first introduce all publications discovered through DBLP which matched the targeted domain (i.e., Argument Quality) and were therefore considered for our analysis. Here, we have removed 9 entries from the initial list of publications, which were either duplicated, preprints (e.g., arXiv entries[8]), or not actual publications (e.g., presentations).

Second, we list the publications found through the initial search with DBLP, yet excluded from the further analysis for at least one of the following reasons: the topic is not related or has a different focus (e.g., uses the argument quality to measure its effect on consumer behavior) or the paper discusses a different type of quality (e.g., argumentation applied to food quality).

Third, we introduce all publications which we collected from the EACL 2024 proceedings. Note that we manually added these for completeness, as the conference venue had just finished at the time of our final paper collection, however none of the EACL 2024 paper had been yet been published on the DBLP platform and would have thus remained unintentionally excluded.

Lastly, we present a list of all publications discovered through Snowballing - i.e., publications, which were referenced within the DBLP corpus, which passed the exclusion process. At this step, preprints were excluded from the list. Note that the scope of the publications discovered through Snowballing is somewhat broader, thus allowing us to expand the surveyed corpus to a total of 211 publications.

[8] arxiv.org

DBLP

Alhamzeh, Alaa (2023). 'Financial Argument Quality Assessment in Earnings Conference Calls'. In: *Database and Expert Systems Applications - 34th International Conference, DEXA 2023, Penang, Malaysia, August 28-30, 2023, Proceedings, Part II.* Ed. by Christine Strauss et al. Vol. 14147. Lecture Notes in Computer Science. Springer, pp. 65–81.

Arnhold, Niclas, Philipp Rösner and Tobias Xylander (2022). 'Quality-Aware Argument Re-Ranking for Comparative Questions'. In: *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022.* Ed. by Guglielmo Faggioli et al. Vol. 3180. CEUR Workshop Proceedings. CEUR-WS.org, pp. 3105–3114.

Belland, Brian R et al. (2017). 'High School Students' Collaboration and Engagement With Scaffolding and Information as Predictors of Argument Quality During Problem-Based Learning'. In: Philadelphia, PA: International Society of the Learning Sciences.

Britner, Sebastian, 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 2023, Birmingham, United Kingdom, October 21-25, 2023.* Ed. by Ingo Frommholz et al. ACM, pp. 5015–5020.

Brudvik, Ole C et al. (2006). 'Assessing the impact of a structured argumentation board on the quality of students' argumentative writing skills'. In: *Frontiers in Artificial Intelligence and Applications* 151, p. 141.

Ceolin, Davide, Giuseppe Primiero, Michael Soprano et al. (2022). 'Transparent assessment of information quality of online reviews using formal argumentation theory'. In: *Information Systems* 110, p. 102107. ISSN: 0306-4379.

Ceolin, Davide, Giuseppe Primiero, Jan Wielemaker et al. (2021). 'Assessing the quality of online reviews using formal argumentation theory'. In: *International Conference on Web Engineering.* Springer, pp. 71–87.

Chimetto, Alessandro et al. (2022). 'SEUPD@CLEF: Team hextech on Argument Retrieval for Comparative Questions. The importance of adjectives in documents quality evaluation'. In: *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022.* Ed. by Guglielmo Faggioli et al. Vol. 3180. CEUR Workshop Proceedings. CEUR-WS.org, pp. 3041–3054.

Clark, Douglas et al. (2007). 'Evaluating the quality of dialogical argumentation in CSCL: moving beyond an analysis of formal structure'. In: *Proceedings of the 8th iternational conference on Computer supported collaborative learning.* International Society of the Learning Sciences, Inc.

Clark, Douglas B. and Victor D. Sampson (2005). 'Analyzing the quality of argumentation supported by personally-seeded discussions'. In: *The Next 10 Years! Proceedings of the 2005 Conference on Computer Support for Collaborative Learning, CSCL '05, Taipei, Taiwan, May 30 - June 4, 2005.* Ed. by Tak-Wai Chan. International Society of the Learning Sciences, pp. 76–85.

Dorton, Stephen L et al. (2021). 'Up for debate: Effects of formal structure on argumentation quality in a crowdsourcing platform'. In: *International Conference on Human-Computer Interaction.* Springer, pp. 36–53.

Dumani, Lorik and Ralf Schenkel (2020a). 'Quality-aware ranking of arguments'. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 335–344.

– (2020b). 'Ranking Arguments by Combining Claim Similarity and Argument Quality Dimensions.' In: *CLEF (Working Notes).*

El Baff, Roxanne et al. (Oct. 2018). 'Challenge or Empower: Revisiting Argumentation Quality in a News Editorial Corpus'. In: *Proceedings of the 22nd Conference on Computational Natural Language Learning.* Brussels, Belgium: Association for Computational Linguistics, pp. 454–464.

Falk, Neele, Iman Jundi et al. (Nov. 2021). 'Predicting Moderation of Deliberative Arguments: Is Argument Quality the Key?' In: *Proceedings of the 8th Workshop on Argument Mining.* Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 133–141.

Falk, Neele and Gabriella Lapesa (May 2023). 'Bridging Argument Quality and Deliberative Quality Annotations with Adapters'. In: *Findings of the Association for Computational Linguistics: EACL 2023.* Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 2469–2488.

Favreau, Charles-Olivier, Amal Zouaq and Sameer Bhatnagar (2022). 'Learning to Rank with BERT for Argument Quality Evaluation'. In: *Proceedings of the Thirty-Fifth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2022, Hutchinson Island, Jensen Beach, Florida, USA, May 15-18, 2022.* Ed. by Roman Barták, Fazel Keshtkar and Michael Franklin.

Fromm, Michael, Max Berrendorf, Evgeniy Faerman et al. (July 2023). ‘Cross-Domain Argument Quality Estimation’. In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers, Jordan Boyd-Graber and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics.

Fromm, Michael, Max Berrendorf, Johanna Reiml et al. (2022). ‘Towards a Holistic View on Argument Quality Prediction’. In: *CoRR* abs/2205.09803.

Gienapp, Lukas (2021). ‘Quality-aware Argument Retrieval with Topical Clustering.’ In: *CLEF (Working Notes)*, pp. 2366–2373.

Gienapp, Lukas et al. (July 2020). ‘Efficient Pairwise Annotation of Argument Quality’. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5772–5781.

Green, Tommaso, Luca Moroldo and Alberto Valente (2021). ‘Exploring BERT Synonyms and Quality Prediction for Argument Retrieval.’ In: *CLEF (Working Notes)*, pp. 2374–2388.

Gretz, Shai et al. (2020). ‘A large-scale dataset for argument quality ranking: Construction and analysis’. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34, pp. 7805–7813.

Guo, Kai et al. (2023). ‘Effects of chatbot-assisted in-class debates on students’ argumentation skills and task motivation’. In: *Computers Education* 203, p. 104862.

Hahn, Ulrike and Jos Hornikx (2016). ‘A normative framework for argument quality: Argumentation schemes with a Bayesian foundation’. In: *Synthese* 193, pp. 1833–1873.

Hasan, Md Kamrul et al. (2021). ‘Hitting your MARQ: Multimodal ARgument quality assessment in long debate video’. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6387–6397.

Hinton, Martin and Jean H. M. Wagemans (2023). ‘How persuasive is AI-generated argumentation? An analysis of the quality of an argumentative text produced by the GPT-3 AI text generator’. In: *Argument Comput.* 14.1, pp. 59–74.

Joshi, Omkar, Priya Pitre and Yashodhara Haribhakta (July 2023). ‘ArgAnalysis35K : A large-scale dataset for Argument Quality Analysis’. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics.

Klebanov, Beata Beigman et al. (2016). ‘Argumentation: Content, structure, and relationship with essay quality’. In: *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pp. 70–75.

Lauscher, Anne et al. (Dec. 2020). ‘Rhetoric, Logic, and Dialectic: Advancing Theory-based Argument Quality Assessment in Natural Language Processing’. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 4563–4574.

Liu, Zhexiong et al. (2023). ‘Predicting the Quality of Revisions in Argumentative Writing’. In: *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@ACL 2023, Toronto, Canada, 13 July 2023*. Ed. by Ekaterina Kochmar et al. Association for Computational Linguistics.

Marro, Santiago, Elena Cabrio and Serena Villata (2022). ‘Graph Embeddings for Argumentation Quality Assessment’. In: *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*. Ed. by Yoav Goldberg, Zornitsa Kozareva and Yue Zhang. Association for Computational Linguistics, pp. 4154–4164.

Meer, Michiel van der et al. (2022). ‘Will It Blend? Mixing Training Paradigms & Prompting for Argument Quality Prediction’. In: *Proceedings of the 9th Workshop on Argument Mining, ArgMining@COLING*. Ed. by Gabriella Lapesa et al.

Min, Chen and Fei Shen (2023). ‘Online incivility, argument quality and public expression in China: Exploring the moderating role of education level and opinion congruency’. In: *Telematics Informatics* 82, p. 102010.

Mun, Y Yi et al. (2013). ‘Untangling the antecedents of initial trust in Web-based health information: The roles of argument quality, source expertise, and user perceptions of information quality and risk’. In: *Decision support systems* 55.1, pp. 284–295.

Ng, Lily et al. (Dec. 2020). ‘Creating a Domain-diverse Corpus for Theory-based Argument Quality Assessment’. In: *Proceedings of the 7th Workshop on Argument Mining*. Online: Association for Computational Linguistics, pp. 117–126.

Nilles, Markus, Lorik Dumani and Ralf Schenkel (2021). ‘QuARk: A GUI for Quality-Aware Ranking of Arguments’. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2546–2549.

Noroozi, Omid et al. (2023). ‘Design, implementation, and evaluation of an online supported peer feedback module to enhance students’ argumentative essay quality’. In: *Educ. Inf. Technol.* 28.10, pp. 12757–12784.

Plenz, Moritz, Raphael Buchmüller and Alexander Bondarenko (2023). ‘Argument Quality Prediction for Ranking Documents’. In: *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*. Ed. by Mohammad Aliannejadi et al. Vol. 3497. CEUR Workshop Proceedings. CEUR-WS.org, pp. 3119–3130.

Rebello, Carina M, Eleanor Sayre and N Sanjay Rebello (2012). ‘Effects of Argumentation Scaffolds and Problem Representation on Students’ Solutions and Argumentation Quality in Physics’. In: *International Society of the Learning Sciences*.

Saveleva, Ekaterina et al. (2021). ‘Graph-based argument quality assessment’. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pp. 1268–1280.

Shiota, Tsukasa and Kazutaka Shimada (2020). ‘The discussion corpus toward argumentation quality assessment in multi-party conversation’. In: *2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI)*. IEEE, pp. 280–283.

Skitalinskaya, Gabriella, Jonas Klaff and Henning Wachsmuth (Apr. 2021). ‘Learning From Revisions: Quality Assessment of Claims in Argumentation at Scale’. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 1718–1729.

Sun, Xichen, Wenhan Chao and Zhunchen Luo (2021). ‘Syntax and Coherence-The Effect on Automatic Argument Quality Assessment’. In: *Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part II 10*. Springer, pp. 3–12.

Toledo, Assaf et al. (Nov. 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 5625–5635.

Wachsmuth, Henning, 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*, pp. 1680–1691.

Wachsmuth, Henning, Nona Naderi, Ivan Habernal et al. (July 2017). ‘Argumentation Quality Assessment: Theory vs. Practice’. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 250–255.

Wachsmuth, Henning, Nona Naderi, Yufang Hou et al. (Apr. 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*. Valencia, Spain: Association for Computational Linguistics, pp. 176–187.

Wachsmuth, Henning and Till Werner (Dec. 2020). ‘Intrinsic Quality Assessment of Arguments’. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 6739–6745.

Wang, Yiran et al. (2023). ‘Contextual Interaction for Argument Post Quality Assessment’. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10420–10432.

DBLP excluded

Ajayi, Oluwakemi and James Zhang (2022). 'The Joint Effects of Argument Quality and Interactivity on Nonprofessional Investors' Perceptions of Disclosure Credibility and Investment Decisions'. In: *J. Inf. Syst.* 36.3, pp. 1–26.

Alhamzeh, Alaa (2023). 'Language reasoning by means of argument mining and argument quality'. PhD thesis. Universität Passau.

Bourguet, Jean-Rémi et al. (2013). 'An artificial intelligence-based approach to deal with argumentation applied to food quality in a public health policy'. In: *Expert Systems with Applications* 40.11, pp. 4539–4546.

Canelas, Jesus Angel Fernandez, Quintin Martin Martin and Juan Manuel Corchado Rodriguez (2013). 'Argumentative sox compliant and quality decision support intelligent expert system over the suppliers selection process'. In: *Applied Computational Intelligence and Soft Computing* 2013, pp. 7–7.

Chen, Xi et al. (2023). 'How to Promote COVID-19 Vaccination in the Digital Media Age: The Persuasive Effects of News Frames and Argument Quality'. In: *Systems* 11.10, p. 491.

Furman, Damián et al. (2023). 'High-quality argumentative information in low resources approaches improve counter-narrative generation'. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2942–2956.

Ha, Sangwook and JoongHo Ahn (2011). 'Why are you sharing others' tweets?: The impact of argument quality and source credibility on information sharing behavior'. In: *ICIS 2011 Proceedings.*

Huhn, Michaela and Axel Zechner (2009). 'Analysing dependability case arguments using quality models'. In: *Computer Safety, Reliability, and Security: 28th International Conference, SAFECOMP 2009, Hamburg, Germany, September 15-18, 2009. Proceedings 28.* Springer, pp. 118–131.

Ishikawa, Fuyuki (2018). 'Concepts in quality assessment for machine learning-from test data to arguments'. In: *Conceptual Modeling: 37th International Conference, ER 2018, Xi'an, China, October 22–25, 2018, Proceedings 37.* Springer, pp. 536–544.

Jin, Gongye (2016). 'High-quality Knowledge Acquisition of Predicate-argument Structures for Syntactic and Semantic Analysis'. In.

Karlsson, Gunnar and Ignacio Más (2007). 'Quality of service and the end-to-end argument'. In: *IEEE Network* 21.6, pp. 16–21.

Ko, Min Jung and Yong Jin Kim (2015). 'Argument Quality or Peripheral Cues: What Makes an Auction Deal Successful in eBay?' In: *Intelligent Decision Technologies: Proceedings of the 7th KES International Conference on Intelligent Decision Technologies (KES-IDT 2015).* Springer, pp. 357–372.

Linden, Dirk van der (2015). 'An argument for more user-centric analysis of modeling languages' visual notation quality'. In: *Advanced Information Systems Engineering Workshops: CAiSE 2015 International Workshops, Stockholm, Sweden, June 8-9, 2015, Proceedings 27.* Springer, pp. 114–120.

Noroozi, Omid, Harm Biemans and Martin Mulder (2016). 'Relations between scripted online peer feedback processes and quality of written argumentative essay'. In: *The Internet and Higher Education* 31, pp. 20–31.

Rach, Niklas et al. (2021). 'Estimating subjective argument quality aspects from social signals in argumentative dialogue systems'. In: *IEEE Access* 9, pp. 11610–11621.

Shin, Soo Yun et al. (2017). 'Investigating moderating roles of goals, reviewer similarity, and self-disclosure on the effect of argument quality of online consumer reviews on attitude formation'. In: *Computers in Human Behavior* 76, pp. 218–226.

Singh, Keshav et al. (Nov. 2021). 'Exploring Methodologies for Collecting High-Quality Implicit Reasoning in Arguments'. In: *Proceedings of the 8th Workshop on Argument Mining.* Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 57–66.

Suarez, Pablo Accuosto (2021). 'Mining arguments in scientific abstracts. Application to argumentative quality assessment'. PhD thesis. Universitat Pompeu Fabra.

Tsai, Pei-Shan (2023). 'Research on information searching strategies in high school students' quality of argumentative essay writing'. In: *Interactive Learning Environments* 31.10, pp. 6799–6817.

Van Snyder, W (1997). *Testing functions of one and two arguments.* Springer.

Wall, Jeffrey D and Merrill Warkentin (2019). 'Perceived argument quality's effect on threat and coping appraisals in fear appeals: An experiment and exploration of realism check heuristics'. In: *Information & Management* 56.8, p. 103157.

Xu, Xi and Zhong Yao (2015). 'Understanding the role of argument quality in the adoption of online reviews: An empirical study integrating value-based decision and needs theory'. In: *Online Information Review* 39.7, pp. 885–902.

EACL24

Falk, Neele et al. (2024). ‘Moderation in the Wild: Investigating User-Driven Moderation in Online Discussions’. In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 992–1013.

Gao, Yingqiang et al. (2024). ‘Evaluating Unsupervised Argument Aligners via Generation of Conclusions of Structured Scientific Abstracts’. In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 151–160.

Tang, An, Xiuzhen Jenny Zhang and Minh Dinh (2024). ‘Aspect-based Key Point Analysis for Quantitative Summarization of Reviews’. In: *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 1419–1433.

Zhang, Jianguo et al. (2024). ‘DialogStudio: Towards Richest and Most Diverse Unified Dataset Collection for Conversational AI’. In: *Findings of the Association for Computational Linguistics: EACL 2024*. Ed. by Yvette Graham and Matthew Purver. St. Julian’s, Malta: Association for Computational Linguistics.

Snowballing

Abbott, Rob et al. (2016). ‘Internet Argument Corpus 2.0: An SQL schema for Dialogic Social Media and the Corpora to go with it’. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. Ed. by Nicoletta Calzolari et al. European Language Resources Association (ELRA).

Afrin, Tazin and Diane J. Litman (2018). ‘Annotation and Classification of Sentence-level Revision Improvement’. In: *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications@NAACL-HLT 2018, New Orleans, LA, USA, June 5, 2018*. Ed. by Joel R. Tetreault et al. Association for Computational Linguistics, pp. 240–246.

Ajjour, Yamen, Henning Wachsmuth, Dora Kiesel et al. (2018). ‘Visualization of the Topic Space of Argument Search Results in args.me’. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*. Ed. by Eduardo Blanco and Wei Lu. Association for Computational Linguistics, pp. 60–65.

Ajjour, Yamen, Henning Wachsmuth, Johannes Kiesel et al. (2019). ‘Data Acquisition for Argument Search: The args.me Corpus’. In: *KI 2019: Advances in Artificial Intelligence*. Ed. by Christoph Benzmüller and Heiner Stuckenschmidt. Cham: Springer International Publishing, pp. 48–59.

Alhindi, Tariq et al. (2022). ‘Multitask Instruction-based Prompting for Fallacy Recognition’. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8172–8187.

Baff, Roxanne El et al. (2018). ‘Challenge or Empower: Revisiting Argumentation Quality in a News Editorial Corpus’. In: *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*. Ed. by Anna Korhonen and Ivan Titov. Association for Computational Linguistics, pp. 454–464.

– (2020). ‘Analyzing the Persuasive Effect of Style in News Editorial Argumentation’. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. Ed. by Dan Jurafsky et al. Association for Computational Linguistics, pp. 3154–3160.

Bar-Haim, Roy, Lilach Eden et al. (2021). ‘Every Bite Is an Experience: Key Point Analysis of Business Reviews’. 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), pp. 3376–3386.

Bar-Haim, Roy, Yoav Kantor et al. (2020). 'Quantitative argument summarization and beyond: Cross-domain key point analysis'. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 39–49.

Basave, Amparo Elizabeth Cano and Yulan He (2016). 'A Study of the Impact of Persuasive Argumentation in Political Debates'. In: *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*. Ed. by Kevin Knight, Ani Nenkova and Owen Rambow. The Association for Computational Linguistics, pp. 1405–1413.

Bentahar, Jamal, Bernard Moulin and Micheline Bélanger (2010). 'A taxonomy of argumentation models used for knowledge representation'. In: *Artif. Intell. Rev.* 33.3, pp. 211–259.

Blanchard, Daniel et al. (2013). 'TOEFL11: A CORPUS OF NON-NATIVE ENGLISH'. In: *ETS Research Report Series* 2013.2, pp. i–15.

Boltuzic, Filip and Jan Snajder (2015). 'Identifying Prominent Arguments in Online Debates Using Semantic Textual Similarity'. In: *Proceedings of the 2nd Workshop on Argumentation Mining, ArgMining@HLT-NAACL 2015, June 4, 2015, Denver, Colorado, USA*. The Association for Computational Linguistics, pp. 110–115.

Boudry, Maarten, Fabio Paglieri and Massimo Pigliucci (2015). 'The Fake, the Flimsy, and the Fallacious: Demarcating Arguments in Real Life'. In: *Argumentation* 29.4, pp. 10–1007.

Bowman, Samuel R. et al. (Sept. 2015). 'A large annotated corpus for learning natural language inference'. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Ed. by Lluís Màrquez, Chris Callison-Burch and Jian Su. Lisbon, Portugal: Association for Computational Linguistics, pp. 632–642.

Braunstain, Liora et al. (2016a). 'Supporting Human Answers for Advice-Seeking Questions in CQA Sites'. In: *Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings*. Ed. by Nicola Ferro et al. Vol. 9626. Lecture Notes in Computer Science. Springer, pp. 129–141.

– (2016b). 'Supporting human answers for advice-seeking questions in CQA sites'. In: *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*. Springer, pp. 129–141.

Cabrio, Elena and Serena Villata (2012). 'Combining Textual Entailment and Argumentation Theory for Supporting Online Debates Interactions'. In: *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 2: Short Papers*. The Association for Computer Linguistics, pp. 208–212.

Chalaguine, Lisa Andreevna and Claudia Schulz (2017). 'Assessing Convincingness of Arguments in Online Debates with Limited Number of Features'. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Student Research Workshop*. Ed. by Florian Kunneman et al. Association for Computational Linguistics, pp. 75–83.

Chen, Xi et al. (2013). 'Pairwise ranking aggregation in a crowdsourced setting'. In: *Sixth ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4-8, 2013*. Ed. by Stefano Leonardi et al. ACM, pp. 193–202.

Correnti, Richard et al. (2013). 'Assessing students' skills at writing analytically in response to texts.' In: *The Elementary School Journal* 144.2, pp. 142–177.

Ding, Yuning, Marie Bexte and Andrea Horbach (2023). 'Score It All Together: A Multi-Task Learning Study on Automatic Scoring of Argumentative Essays'. In: *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*. Ed. by Anna Rogers, Jordan L. Boyd-Graber and Naoaki Okazaki. Association for Computational Linguistics, pp. 13052–13063.

Dondio, Pierpaolo (2014). 'Towards a Computational Analysis of Probabilistic Argumentation Frameworks'. In: *Cybernetics and Systems* 45.3, pp. 254–278.

Dumani, Lorik, Patrick J. Neumann and Ralf Schenkel (2020). 'A Framework for Argument Retrieval - Ranking Argument Clusters by Frequency and Specificity'. In: *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I*. Ed. by Joemon M. Jose et al. Vol. 12035. Lecture Notes in Computer Science. Springer, pp. 431–445.

Dumani, Lorik and Ralf Schenkel (2019). 'A Systematic Comparison of Methods for Finding Good Premises for Claims'. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*. Ed. by Benjamin Piwowarski et al. ACM, pp. 957–960.

Durmus, Esin 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*. Vol. 1.

– (2019). ‘A Corpus for Modeling User and Language Effects in Argumentation on Online Debating’. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 602–607.

Durmus, Esin, Faisal Ladhak and Claire Cardie (2019). ‘The Role of Pragmatic and Discourse Context in Determining Argument Impact’. 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 2019, Hong Kong, China, November 3-7, 2019*. Ed. by Kentaro Inui et al. Association for Computational Linguistics, pp. 5667–5677.

Dursun, Ahmet and Zhi Li (2021). ‘A Systematic Review of Argument-Based Validation Studies in the Field of Language Testing (2000–2018)’. In: *Validity Argument in Language Testing: Case Studies of Validation Research*. Cambridge Applied Linguistics. Cambridge University Press, pp. 45–70.

Egawa, Ryo, Gaku Morio and Katsuhide Fujita (2019). ‘Annotating and analyzing semantic role of elementary units and relations in online persuasive arguments’. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 422–428.

Feng, Vanessa Wei, Ziheng Lin and Graeme Hirst (2014). ‘The Impact of Deep Hierarchical Discourse Structures in the Evaluation of Text Coherence’. In: *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*. Ed. by Jan Hajic and Junichi Tsujii. ACL, pp. 940–949.

Fung, Michelle et al. (2015). ‘ROC speak: semi-automated personalized feedback on nonverbal behavior from recorded videos’. In: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp 2015, Osaka, Japan, September 7-11, 2015*. Ed. by Kenji Mase et al. ACM, pp. 1167–1178.

Ghosh, Debanjan et al. (2016). ‘Coarse-grained argumentation features for scoring persuasive essays’. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 549–554.

Gleize, Martin et al. (2019). ‘Are You Convinced? Choosing the More Convincing Evidence with a Siamese Network’. In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28-August 2, 2019, Volume 1: Long Papers*. Ed. by Anna Korhonen, David R. Traum and Lluís Màrquez. Association for Computational Linguistics, pp. 967–976.

Goffredo, Pierpaolo et al. (2022). ‘Fallacious Argument Classification in Political Debates’. In: *Thirty-First International Joint Conference on Artificial Intelligence {IJCAI-22}*, pp. 4143–4149.

Granger, Sylviane et al. (2009). *International Corpus of Learner English Vol. 2*. Louvain-la-Neuve: Presses universitaires de Louvain.

Gurcke, Timon, Milad Alshomary and Henning Wachsmuth (2021). ‘Assessing the Sufficiency of Arguments through Conclusion Generation’. In: *Proceedings of the 8th Workshop on Argument Mining, ArgMining@EMNLP 2021, Punta Cana, Dominican Republic, November 10-11, 2021*. Ed. by Khalid Al Khatib, Yufang Hou and Manfred Stede. Association for Computational Linguistics, pp. 67–77.

Habernal, Ivan 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, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. Ed. by Jian Su, Xavier Carreras and Kevin Duh. The Association for Computational Linguistics, pp. 1214–1223.

– (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, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

– (2017). ‘Argumentation Mining in User-Generated Web Discourse’. In: *Comput. Linguistics* 43.1, pp. 125–179.

Habernal, Ivan, Raffael Hannemann et al. (2017). ‘Argotario: Computational Argumentation Meets Serious Games’. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 7–12.

Habernal, Ivan, Henning Wachsmuth et al. (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)*, pp. 386–396.

Hasan, Kazi Saidul 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 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. Ed. by Alessandro Moschitti, Bo Pang and Walter Daelemans. ACL, pp. 751–762.

Heinisch, Philipp et al. (2023). 'ACCEPT at SemEval-2023 Task 3: An Ensemble-based Approach to Multilingual Framing Detection'. In: *Proceedings of the The 17th International Workshop on Semantic Evaluation, SemEval@ACL 2023, Toronto, Canada, 13-14 July 2023*. Ed. by Atul Kr. Ojha et al. Association for Computational Linguistics, pp. 1347–1358.

Hidey, Christopher et al. (2017). 'Analyzing the semantic types of claims and premises in an online persuasive forum'. In: *Proceedings of the 4th Workshop on Argument Mining*. Columbia Univ., New York, NY (United States).

Hovy, Dirk et al. (June 2013). 'Learning Whom to Trust with MACE'. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Lucy Vanderwende, Hal Daumé III and Katrin Kirchhoff. Atlanta, Georgia: Association for Computational Linguistics, pp. 1120–1130.

Hulpus, Ioana et al. (2019). 'Towards Explaining Natural Language Arguments with Background Knowledge'. In: *Joint Proceedings of the 6th International Workshop on Dataset PROFlLing and Search & the 1st Workshop on Semantic Explainability co-located with the 18th International Semantic Web Conference (ISWC 2019), Auckland, New Zealand, October 27, 2019*. Ed. by Elena Demidova et al. Vol. 2465. CEUR Workshop Proceedings. CEUR-WS.org, pp. 62–77.

Jin, Zhijing et al. (2022). 'Logical Fallacy Detection'. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 7180–7198.

Kees, Nataliia et al. (2021). 'Active Learning for Argument Strength Estimation'. In: *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pp. 144–150.

Khatib, Khalid Al, Henning Wachsmuth, Matthias Hagen et al. (2017). 'Patterns of Argumentation Strategies across Topics'. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. Ed. by Martha Palmer, Rebecca Hwa and Sebastian Riedel. Association for Computational Linguistics, pp. 1351–1357.

Khatib, Khalid Al, Henning Wachsmuth, Johannes Kiesel et al. (2016). 'A News Editorial Corpus for Mining Argumentation Strategies'. In: *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*. Ed. by Nicoletta Calzolari, Yuji Matsumoto and Rashmi Prasad. ACL, pp. 3433–3443.

Kim, Hyunwoo et al. (2023). 'SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization'. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12930–12949.

Klebanov, Beata Beigman et al. (2016). 'Argumentation: Content, Structure, and Relationship with Essay Quality'. In: *Proceedings of the Third Workshop on Argument Mining, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2016, August 12, Berlin, Germany*. The Association for Computer Linguistics.

Lauscher, Anne et al. (2022). 'Scientia Potentia Est - On the Role of Knowledge in Computational Argumentation'. In: *Trans. Assoc. Comput. Linguistics* 10, pp. 1392–1422.

Levy, Ran et al. (2014). 'Context Dependent Claim Detection'. In: *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*. Ed. by Jan Hajic and Junichi Tsujii. ACL, pp. 1489–1500.

Li, Jialu, Esin Durmus and Claire Cardie (2020). 'Exploring the Role of Argument Structure in Online Debate Persuasion'. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8905–8912.

Liu, Haijing et al. (2017). 'Using Argument-based Features to Predict and Analyse Review Helpfulness'. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. Ed. by Martha Palmer, Rebecca Hwa and Sebastian Riedel. Association for Computational Linguistics, pp. 1358–1363.

Liu, Xin et al. (2021). 'Exploring Discourse Structures for Argument Impact Classification'. In: *ACL-IJCNLP 2021-59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, p. 3958.

Longpre, Liane, Esin Durmus and Claire Cardie (2019a). 'Persuasion of the Undecided: Language vs. the Listener'. In: *Proceedings of the 6th Workshop on Argument Mining, ArgMining@ACL 2019, Florence, Italy, August 1, 2019*. Ed. by Benno Stein and Henning Wachsmuth. Association for Computational Linguistics, pp. 167–176.

Longpre, Liane, Esin Durmus and Claire Cardie (2019b). ‘Persuasion of the Undecided: Language vs. the Listener.’ In: *Proceedings of the 6th Workshop on Argument Mining.*

Lukin, Stephanie M. et al. (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, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers.* Ed. by Mirella Lapata, Phil Blunsom and Alexander Koller. Association for Computational Linguistics, pp. 742–753.

Luu, Kelvin, Chenhao Tan and Noah A Smith (2019). ‘Measuring online debaters’ persuasive skill from text over time’. In: *Transactions of the Association for Computational Linguistics* 7, pp. 537–550.

Morio, Gaku, Ryo Egawa and Katsuhide Fujita (2019). ‘Revealing and predicting online persuasion strategy with elementary units’. 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)*, pp. 6274–6279.

Olshefski, Christopher et al. (2020). ‘The Discussion Tracker Corpus of Collaborative Argumentation’. In: *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020.* Ed. by Nicoletta Calzolari et al. European Language Resources Association, pp. 1033–1043.

Park, Joonsuk, Cheryl Blake and Claire Cardie (2015). ‘Toward machine-assisted participation in eRulemaking: an argumentation model of evaluability’. In: *Proceedings of the 15th International Conference on Artificial Intelligence and Law, ICAIL 2015, San Diego, CA, USA, June 8-12, 2015.* Ed. by Ted Sichelman and Katie Atkinson. ACM, pp. 206–210.

Park, Joonsuk, Sally Klingel et al. (2012). ‘Facilitative moderation for online participation in eRulemaking’. In: *Proceedings of the 13th Annual International Conference on Digital Government Research*, pp. 173–182.

Passonneau, Rebecca J. and Bob Carpenter (2014). ‘The Benefits of a Model of Annotation’. In: *Trans. Assoc. Comput. Linguistics* 2, pp. 311–326.

Peldszus, Andreas and Manfred Stede (2015a). ‘An annotated corpus of argumentative microtexts’. In: *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon.* Vol. 2, pp. 801–815.

– (2015b). ‘Joint prediction in MST-style discourse parsing for argumentation mining’. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015.* Ed. by Lluís Màrquez et al. The Association for Computational Linguistics, pp. 938–948.

Persing, Isaac, Alan Davis and Vincent Ng (2010). ‘Modeling Organization in Student Essays’. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL.* ACL, pp. 229–239.

Persing, Isaac and Vincent Ng (2013). ‘Modeling Thesis Clarity in Student Essays’. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers.* The Association for Computer Linguistics, pp. 260–269.

– (2014). ‘Modeling Prompt Adherence in Student Essays’. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers.* The Association for Computer Linguistics, pp. 1534–1543.

– (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 of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers.* The Association for Computer Linguistics, pp. 543–552.

– (2017a). ‘Lightly-supervised modeling of argument persuasiveness’. In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 594–604.

– (2017b). ‘Why Can’t You Convince Me? Modeling Weaknesses in Unpersuasive Arguments’. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017.* Ed. by Carles Sierra. ijcai.org, pp. 4082–4088.

Plank, Barbara (2022). ‘The ”Problem” of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation’. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022.* Ed. by Yoav Goldberg, Zornitsa Kozareva and Yue Zhang. Association for Computational Linguistics, pp. 10671–10682.

Plenz, Moritz et al. (2023). ‘Similarity-weighted Construction of Contextualized Commonsense Knowledge Graphs for Knowledge-intense Argumentation Tasks’. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*. Ed. by Anna Rogers, Jordan L. Boyd-Graber and Naoaki Okazaki. Association for Computational Linguistics, pp. 6130–6158.

Potash, Peter, 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*. Ed. by Greg Kondrak and Taro Watanabe. Asian Federation of Natural Language Processing, pp. 342–351.

Potash, Peter, Adam Ferguson and Timothy J. Hazen (2019). ‘Ranking Passages for Argument Convincingness’. In: *Proceedings of the 6th Workshop on Argument Mining, ArgMining@ACL 2019, Florence, Italy, August 1, 2019*. Ed. by Benno Stein and Henning Wachsmuth. Association for Computational Linguistics, pp. 146–155.

Potash, Peter and Anna Rumshisky (2017). ‘Towards Debate Automation: a Recurrent Model for Predicting Debate Winners’. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. Ed. by Martha Palmer, Rebecca Hwa and Sebastian Riedel. Association for Computational Linguistics, pp. 2465–2475.

Potthast, Martin et al. (2019). ‘Argument Search: Assessing Argument Relevance’. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*. Ed. by Benjamin Piwowarski et al. ACM, pp. 1117–1120.

Rach, Niklas et al. (2020). ‘Evaluation of Argument Search Approaches in the Context of Argumentative Dialogue Systems’. In: *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*. Ed. by Nicoletta Calzolari et al. European Language Resources Association, pp. 513–522.

Rahimi, Zahra, Diane J. Litman, Richard Correnti et al. (2014). ‘Automatic Scoring of an Analytical Response-To-Text Assessment’. In: *Intelligent Tutoring Systems - 12th International Conference, ITS 2014, Honolulu, HI, USA, June 5-9, 2014. Proceedings*. Ed. by Stefan Trausan-Matu et al. Vol. 8474. Lecture Notes in Computer Science. Springer, pp. 601–610.

Rahimi, Zahra, Diane J. Litman, Elaine Wang et al. (2015). ‘Incorporating Coherence of Topics as a Criterion in Automatic Response-to-Text Assessment of the Organization of Writing’. In: *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, BEA@NAACL-HLT 2015, June 4, 2015, Denver, Colorado, USA*. Ed. by Joel R. Tetreault, Jill Burstein and Claudia Leacock. The Association for Computer Linguistics, pp. 20–30.

Ramesh, Dadi and Suresh Kumar Sanampudi (2021). ‘An automated essay scoring systems: a systematic literature review’. In: *Artificial Intelligence Review* 55, pp. 2495–2527.

Reed, Chris et al. (2008). ‘Language Resources for Studying Argument’. In: *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. European Language Resources Association.

Reimers, Nils et al. (2019). ‘Classification and Clustering of Arguments with Contextualized Word Embeddings’. In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28-August 2, 2019, Volume 1: Long Papers*. Ed. by Anna Korhonen, David R. Traum and Lluís Màrquez. Association for Computational Linguistics, pp. 567–578.

Rinott, Ruty et al. (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 2015, Lisbon, Portugal, September 17-21, 2015*. Ed. by Lluís Màrquez et al. The Association for Computational Linguistics, pp. 440–450.

Romberg, Julia, Laura Mark and Tobias Escher (2022). ‘A Corpus of German Citizen Contributions in Mobility Planning: Supporting Evaluation Through Multidimensional Classification’. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*. Ed. by Nicoletta Calzolari et al. European Language Resources Association, pp. 2874–2883.

Rosenfeld, Ariel and Sarit Kraus (2015). ‘Providing Arguments in Discussions Based on the Prediction of Human Argumentative Behavior’. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*. Ed. by Blai Bonet and Sven Koenig. AAAI Press, pp. 1320–1327.

– (2016). ‘Providing Arguments in Discussions on the Basis of the Prediction of Human Argumentative Behavior'. In: *ACM Trans. Interact. Intell. Syst.* 6.4.

Segura-Tinoco, Andrés and Iván Cantador (2023). 'ARGAEL: ARGument Annotation and Evaluation tooL'. In: *SoftwareX* 23, p. 101410.

Shmueli-Scheuer, Michal et al. (2019). 'Detecting persuasive arguments based on author-reader personality traits and their interaction'. In: *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, pp. 211–215.

Shnarch, Eyal et al. (2018). 'Will it Blend? Blending Weak and Strong Labeled Data in a Neural Network for Argumentation Mining'. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers.* Ed. by Iryna Gurevych and Yusuke Miyao. Association for Computational Linguistics, pp. 599–605.

Simpson, Edwin and Iryna Gurevych (2018). 'Finding Convincing Arguments Using Scalable Bayesian Preference Learning'. In: *Trans. Assoc. Comput. Linguistics* 6, pp. 357–371.

Skitalinskaya, Gabriella and Henning Wachsmuth (2023). 'To Revise or Not to Revise: Learning to Detect Improvable Claims for Argumentative Writing Support'. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023.* Ed. by Anna Rogers, Jordan L. Boyd-Graber and Naoaki Okazaki. Association for Computational Linguistics, pp. 15799–15816.

Song, Yi et al. (June 2014). 'Applying Argumentation Schemes for Essay Scoring'. In: *Proceedings of the First Workshop on Argumentation Mining.* Ed. by Nancy Green et al. Baltimore, Maryland: Association for Computational Linguistics, pp. 69–78.

Speer, Robyn, Joshua Chin and Catherine Havasi (2017). 'ConceptNet 5.5: An Open Multilingual Graph of General Knowledge'. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.* Ed. by Satinder Singh and Shaul Markovitch. AAAI Press, pp. 4444–4451.

Stab, Christian, Johannes Daxenberger et al. (2018). 'ArgumenText: Searching for Arguments in Heterogeneous Sources'. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 2-4, 2018, Demonstrations.* Ed. by Yang Liu, Tim Paek and Manasi Patwardhan. Association for Computational Linguistics, pp. 21–25.

Stab, Christian and Iryna Gurevych (2014a). 'Annotating Argument Components and Relations in Persuasive Essays'. In: *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland.* Ed. by Jan Hajic and Junichi Tsujii. ACL, pp. 1501–1510.

– (2014b). 'Identifying Argumentative Discourse Structures in Persuasive Essays'. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL.* Ed. by Alessandro Moschitti, Bo Pang and Walter Daelemans. ACL, pp. 46–56.

– (2016). 'Recognizing the Absence of Opposing Arguments in Persuasive Essays'. In: *Proceedings of the Third Workshop on Argument Mining, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2016, August 12, Berlin, Germany.* The Association for Computer Linguistics.

– (2017a). 'Parsing Argumentation Structures in Persuasive Essays'. In: *Comput. Linguistics* 43.3, pp. 619–659.

– (2017b). 'Recognizing Insufficiently Supported Arguments in Argumentative Essays'. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers.* Ed. by Mirella Lapata, Phil Blunsom and Alexander Koller. Association for Computational Linguistics, pp. 980–990.

Stapleton, Paul and Yanming (Amy) Wu (2015). 'Assessing the quality of arguments in students' persuasive writing: A case study analyzing the relationship between surface structure and substance'. In: *Journal of English for Academic Purposes* 17, pp. 12–23.

Swanson, Reid, Brian Ecker and Marilyn A. Walker (2015). 'Argument Mining: Extracting Arguments from Online Dialogue'. In: *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic.* The Association for Computer Linguistics, pp. 217–226.

Tan, Chenhao, Vlad Niculae, Cristian Danescu-Niculescu-Mizil et al. (2016a). 'Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions'. In: *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016.* Ed. by Jacqueline Bourdeau et al. ACM, pp. 613–624.

– (2016b). 'Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith

Online Discussions'. In: *Proceedings of the 25th International Conference on World Wide Web.* WWW '16. Montréal, Québec, Canada: International World Wide Web Conferences Steering Committee, pp. 613–624. ISBN: 9781450341431.

Wachsmuth, Henning, Martin Potthast et al. (2017). 'Building an Argument Search Engine for the Web'. In: *Proceedings of the 4th Workshop on Argument Mining, ArgMining@EMNLP 2017, Copenhagen, Denmark, September 8, 2017.* Ed. by Ivan Habernal et al. Association for Computational Linguistics, pp. 49–59.

Wachsmuth, Henning, Benno Stein and Yamen Ajjour (2017). '"PageRank" for Argument Relevance'. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers.* Ed. by Mirella Lapata, Phil Blunsom and Alexander Koller. Association for Computational Linguistics, pp. 1117–1127.

Wachsmuth, Henning, Martin Trenkmann, Benno Stein and Gregor Engels (2014). 'Modeling Review Argumentation for Robust Sentiment Analysis'. In: *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland.* Ed. by Jan Hajic and Junichi Tsujii. ACL, pp. 553–564.

Wachsmuth, Henning, Martin Trenkmann, Benno Stein, Gregor Engels and Tsvetomira Palakarska (2014). 'A Review Corpus for Argumentation Analysis'. In: *Computational Linguistics and Intelligent Text Processing - 15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6-12, 2014, Proceedings, Part II.* Ed. by Alexander F. Gelbukh. Vol. 8404. Lecture Notes in Computer Science. Springer, pp. 115–127.

Walker, Marilyn A. et al. (2012). 'A Corpus for Research on Deliberation and Debate'. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012.* Ed. by Nicoletta Calzolari et al. European Language Resources Association (ELRA), pp. 812–817.

Wambsganss, Thiemo and Christina Niklaus (2022). 'Modeling Persuasive Discourse to Adaptively Support Students' Argumentative Writing'. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022.* Ed. by Smaranda Muresan, Preslav Nakov and Aline Villavicencio. Association for Computational Linguistics, pp. 8748–8760.

Wang, Lu et al. (2017). 'Winning on the Merits: The Joint Effects of Content and Style on Debate Outcomes'. In: *Trans. Assoc. Comput. Linguistics* 5, pp. 219–232.

Wei, Zhongyu, Yang Liu and Yi Li (Aug. 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).* Ed. by Katrin Erk and Noah A. Smith. Berlin, Germany: Association for Computational Linguistics, pp. 195–200.

Wei, Zhongyu, Yandi Xia et al. (Aug. 2016). 'A Preliminary Study of Disputation Behavior in Online Debating Forum'. In: *Proceedings of the Third Workshop on Argument Mining (ArgMining2016).* Ed. by Chris Reed. Berlin, Germany: Association for Computational Linguistics, pp. 166–171.

Williams, Adina, Nikita Nangia and Samuel R. Bowman (2018). 'A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference'. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers).* Ed. by Marilyn A. Walker, Heng Ji and Amanda Stent. Association for Computational Linguistics, pp. 1112–1122.

Yang, Wonsuk et al. (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 2019, Hong Kong, China, November 3-7, 2019.* Ed. by Kentaro Inui et al. Association for Computational Linguistics, pp. 2954–2963.

Ye, P, UMD EDU and David Doermann (2013). 'Combining preference and absolute judgements in a crowd-sourced setting'. In: *ICML Workshop.*

Zhang, Fan, Homa B. Hashemi et al. (2017). 'A Corpus of Annotated Revisions for Studying Argumentative Writing'. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers.* Ed. by Regina Barzilay and Min-Yen Kan. Association for Computational Linguistics, pp. 1568–1578.

Zhang, Fan, Rebecca Hwa et al. (2016). 'ArgRewrite: A Web-based Revision Assistant for Argumentative Writings'. In: *Proceedings of the Demonstrations Session, NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17,* *2016*. The Association for Computational Linguistics, pp. 37–41.

Zhang, Fan and Diane J. Litman (2015). 'Annotation and Classification of Argumentative Writing Revisions'. In: *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, BEA@NAACL-HLT 2015, June 4, 2015, Denver, Colorado, USA*. Ed. by Joel R. Tetreault, Jill Burstein and Claudia Leacock. The Association for Computer Linguistics, pp. 133–143.

Zhang, Justine et al. (2016). 'Conversational Flow in Oxford-style Debates'. In: *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*. Ed. by Kevin Knight, Ani Nenkova and Owen Rambow. The Association for Computational Linguistics, pp. 136–141.

Ziegenbein, Timon et al. (2023). 'Modeling Appropriate Language in Argumentation'. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*. Ed. by Anna Rogers, Jordan L. Boyd-Graber and Naoaki Okazaki. Association for Computational Linguistics, pp. 4344–4363.