AQuA – Combining Experts' and Non-Experts' Views To Assess Deliberation Quality in Online Discussions Using LLMs

Maike Behrendt[1], Stefan Sylvius Wagner[1], Marc Ziegele[1],
Lena Wilms[1], Anke Stoll[2], Dominique Heinbach[3], Stefan Harmeling[4]
[1]Heinrich Heine University Düsseldorf, Germany [2]Technical University Ilmenau, Germany
[3] Johannes Gutenberg University Mainz, Germany [4]Technical University Dortmund, Germany
[1]{maike.behrendt, stefan.wagner, lena.wilms, marc.ziegele}@uni-duesseldorf.de
[2]anke.stoll@tu-ilmenau.de [3]dominique.heinbach@uni-mainz.de [4]stefan.harmeling@tu-dortmund.de

Abstract

Measuring the quality of contributions in political online discussions is crucial in deliberation research and computer science. Research has identified various indicators to assess online discussion quality, and with deep learning advancements, automating these measures has become feasible. While some studies focus on analyzing specific quality indicators, a comprehensive quality score incorporating various deliberative aspects is often preferred. In this work, we introduce AQuA, an additive score that calculates a unified deliberative quality score from multiple indices for each discussion post. Unlike other singular scores, AQuA preserves information on the deliberative aspects present in comments, enhancing model transparency. We develop adapter models for 20 deliberative indices, and calculate correlation coefficients between experts' annotations and the perceived deliberativeness by non-experts to weigh the individual indices into a single deliberative score. We demonstrate that the AQuA score can be computed easily from pre-trained adapters and aligns well with annotations on other datasets that have not be seen during training. The analysis of experts' vs. non-experts' annotations confirms theoretical findings in the social science literature.

Keywords: deliberative quality, adapter models, quality score

1. Introduction

In the evolving landscape of democratic discourse, the concept of deliberation stands as a cornerstone, embodying the exchange of ideas, critical discussion, and consensus-building among citizens (Dryzek, 2002). Central to the efficacy of these deliberations is their quality, a multifaceted construct traditionally gauged by dimensions such as rationality, civility, reciprocity, and constructiveness (Friess and Eilders, 2015). More recent research has explored various indicators of deliberative quality in online discussions (Steenbergen et al., 2003; Friess and Eilders, 2015; Scudder, 2022). However, most of these approaches require manual annotation of discussion data from trained coders and serve to analyze the discussion in retrospect. As the digital age drives an increasing volume of public conversations onto online platforms, the demand to assess their quality through the previously mentioned dimensions in an automated, scalable manner is growing (Diakopoulos, 2015; Beauchamp, 2020).

Previous efforts have demonstrated the potential of using natural language processing (NLP) and machine learning algorithms to automatically identify features of deliberation such as argumentative structure, emotional tone, and engagement patterns (Lawrence and Reed, 2020; Acheampong et al., 2020; Shin and Rask, 2021). The interest in automating such assessments, with projects like the one implemented by Falk and Lapesa (2023a) in their examination of argument and deliberative quality with adapter models (Houlsby et al., 2019), is growing.

Motivated by this research, this study introduces AQuA, an index to measure the deliberative quality of individual comments in online discussions with a single score. While there is an ongoing debate on the usefulness of aggregating multiple indices of deliberation (Bächtiger et al., 2022), we argue that for some tasks a single value, composed of several theoretically based criteria is favorable. Our approach combines predictions on various dimensions of deliberation with insights gained from both expert and non-expert evaluations, resulting in a single deliberative quality score. We make use of data that has been annotated from both trained experts and crowd annotators, representing the non-experts' view. We calculate correlation coefficients between the annotated deliberative quality criteria and the perceived deliberativeness of the comments to attribute importance to each individual criterion.

Our contributions:

1. We train 20 adapter models on aspects of deliberation to form the basis for a single deliberation score.

2. To combine the automated predictions in a meaningful way, we calculate the correlation

coefficients between experts' and non-experts' assessments of deliberative quality.

3. We define a single normalized score using the correlations as weights, hereby, creating an interpretable and explainable measure for deliberative quality.

4. Finally, we show in experiments that our score can automatically assess the deliberative quality of discussion comments.

Our method consists of two components: (1) the utilization of adapters trained on discrete facets of deliberation, and (2) the integration of correlations between annotations from experts and non-experts to establish a normalized score for deliberative quality. In developing this index, we extensively test and evaluate its effectiveness across diverse datasets, demonstrating its utility in real-world applications. By doing so, we aim to contribute to the burgeoning field of computational social science, offering scholars, policymakers, and practitioners a tool to monitor and analyze public dialogues. Our trained adapter weights and the code for calculating AQuA scores are available under `https://github.com/mabehrendt/AQuA`.

2. Related Work

Before explaining our approach in detail, we give an overview on the previous work to quantify aspects of deliberation in online discussions and the adapter approach to efficiently train language models for downstream tasks.

2.1. Deliberative Quality Indices

Various attempts have been made in the literature to conceptualize deliberation aspects to assess the quality of discourse. Here, we provide a summary of key indicators and metrics proposed in this domain.

The *Deliberative Quality Index* (DQI), introduced by Steenbergen et al. (2003) and further refined by Bächtiger et al. (2022), is a prominent and frequently applied metric for evaluating deliberative quality. The DQI comprises five dimensions: *equality of participation*, *level of justification*, *content of justification*, *respect*, and *constructive politics*. These dimensions are assessed for each contribution and averaged for a single speaker.

Scudder's (2022) *Listening Quality Index* (LQI) emphasizes deliberative listening as a crucial factor in communication quality, organizing elements of existing measures into a hierarchical scale. This scale progresses from minimal listening to a stage where the speaker feels acknowledged, emphasizing the sequential fulfillment of criteria. The LQI differentiates between speakers and listeners, considering not just the contributions to the dialogue but also the participants' behavior and their feeling of being heard.

The *Deliberative Reason Index* (DRI) by Niemeyer et al. (2024) seeks to capture deliberative quality at the group reasoning level rather than evaluating individual contributions. This approach, akin to the LQI, employs surveys conducted before and after discussions to gauge participants' views and preferences on debated topics, calculating agreement scores that are then aggregated to a group score.

Although referred to as indices, the discussed methodologies do not necessarily provide a single index. They often yield multiple metrics rather than a singular measure, demanding a comprehensive evaluation to determine the overall quality of contributions or debates. Friess et al. (2021) suggest aggregating the presence of deliberative qualities — rationality, respect, reciprocity, and civility — and computing their average to establish a quality ratio, treating each criterion with equal importance. We argue, however, that certain aspects may be more important than others to estimate the deliberative quality of a contribution (Chen, 2017).

While the indices presented are valuable for in-depth political debate analysis, their application requires extensive effort from trained coders for annotation and reliability assessments. To streamline the analysis of the deliberative quality of online discussions, several automation proposals have emerged. For instance, Wyss et al. (2015) employ cognitive complexity to analyze Swiss parliamentary debates, using indicators derived from the Linguistic Inquiry and Word Count (LIWC) dictionary (Tausczik and Pennebaker, 2010). Gold et al. (2015) automate the measurement and annotation of features like participation and justification, subsequently employing a visual analytics system for data representation. Fournier-Tombs and Di Marzo Serugendo (2020) introduced DelibAnalysis, a framework for predicting the DQI of online discussion contributions through machine learning, while Shin and Rask (2021) proposed leveraging network and time-series analyzes to assess deliberation criteria automatically.

Our proposed method seeks to bridge the gap between NLP techniques and the theoretical aspects of deliberative quality assessment. We introduce the AQuA score to (i) combine the theoretical underpinnings of deliberation with the comment quality in online debates as perceived by non-experts, and thereby (ii) offering a tool to quantify deliberation aspects through advanced deep learning methods.

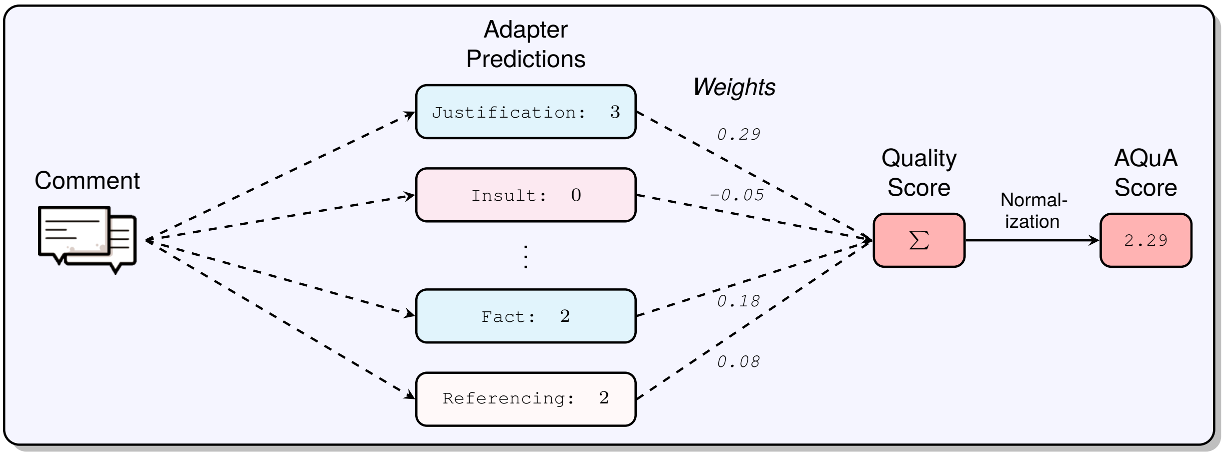

Figure 1: AQuA calculates a single score for deliberativeness from weighted adapter predictions on 20 different deliberative aspects. The adapter predictions are weighted by the correlation coefficients between each deliberative aspect and the perception of crowd workers about whether a comment is deliberative or not. The normalized score can then be used to compare the deliberative quality of individual comments.

2.2. Adapters

Adapters, as introduced by Rebuffi et al. (2017) are an efficient approach to customize pre-trained language models like RoBERTa (Liu et al., 2019) for specific tasks. This method involves the integration of additional bottleneck layers into the model for each distinct task, which adds new weights while leaving the original pre-trained weights unaltered.

The concept of adapter layers was first applied to NLP by Houlsby et al. (2019), who adapted the Transformer architecture (Vaswani et al., 2017) to include these layers. The design of the adapter involves compressing the input's dimensionality to a significantly smaller size, applying a non-linear function, and incorporating a skip-connection to circumvent the bottleneck, with task-specific layer normalization parameters also being adjustable.

The strategic insertion of adapter layers has been a focus of research, with Houlsby et al. (2019) positioning them subsequent to both the multi-head attention and feed-forward layers within the Transformer architecture. Pfeiffer et al. (2021) found in an extensive search on architectural parameters, that placing only one adapter after the feed forward layer in the Transformer works best throughout all their experiments. We also apply this architecture for our models. The introduction of AdapterHub by Pfeiffer et al. (2020) and the adapters library by Poth et al. (2023) further facilitated the sharing and reuse of pre-trained adapters within the community.

Subsequent studies, such as those by Mendonca et al. (2022), explored the training of individual adapters for dialogue quality estimation, and the use of AdapterFusion (Pfeiffer et al., 2021) to merge features from different adapters. Falk and Lapesa (2023a) trained 20 adapters on features for argument and deliberative quality to examine their dependencies. In our work, we follow a similar path to train adapters to evaluate specific aspects of deliberative quality and subsequently combine them using correlation coefficients between experts' and non-experts' annotations, to create a single deliberative quality metric.

3. AQuA: An Additive Score for Deliberative Quality

With AQuA we propose a metric for assessing the quality of individual comments in online discussions. Our approach combines predictions on various dimensions of deliberation with insights gained from both experts' and non-experts' evaluations, resulting in a single deliberative quality score. Our methodology consists of two components: (1) the utilization of adapters trained on discrete facets of deliberation, and (2) the integration of correlations between experts' and non-experts' annotations to establish a normalized score for deliberative quality. We therefore harness annotations of the same data, once labeled by trained experts for a variety of deliberative qualities, such as the degree of justification, and once labeled by non-experts on their personal assessment of the deliberativeness of a comment. We calculate correlation coefficients between each individual deliberative criterion (experts' labels) and the binary indicator for deliberativeness (non-experts' labels).

The idea of our approach is to aggregate individual scores calculated by adapters in a meaningful way to obtain a single score for each comment, in which some aspects contribute more to the perceived deliberativeness than others. For this reason we call our approach AQuA, an "Additive deliberative Quality score with Adapters".

3.1. Datasets

Our analysis is based on three datasets:

1. The KODIE dataset, comprising 13,587 comments that were collected and annotated as part of a scientific study that explored the impact of news organizations' interactive moderation on the deliberative quality of users' political discussions (Heinbach et al., 2022). The comments were posted on the Facebook pages of four German national and regional news outlets with high outreach and diverse audiences. These news outlets delivered data that included all published and deleted/hidden posts and comments on their Facebook pages for a period of 12 weeks per news outlet.
2. The #meinfernsehen2021 (German for my television) dataset (Gerlach and Eilders, 2022) is the result of a large scale citizen participation on the future of public television in Germany. Overall, 1,714 comments from the participation process have been manually coded as part of a quantitative content analysis to examine the discussion quality.
3. The CrowdAnno project Wilms et al. (2023) collected a non-expert representation of deliberative quality via crowd annotations for a subset of, i.a., both the KODIE and #meinfernsehen datasets.

The annotations from two different perspectives are explained in the following.

3.1.1. KODIE & #meinfernsehen - the Experts' View

The KODIE annotation framework (Heinbach et al., 2022), assigns 23 score-based deliberative and further labels on other aspects to each comment. These annotations were conducted by trained coders with a scientific background, focusing on deliberative criteria such as fact claims, relevance to the discussion topic, and respectful engagement with other users. The deliberative criteria can each be assigned to one of the three main dimensions of deliberation (Bächtiger et al., 2009; Esau et al., 2021; Graham, 2010; Coe et al., 2014; Papacharissi, 2004):

Rationality, measured by indicators such as reasoning, solution proposals, and provision of additional knowledge.

Reciprocity, measured as mutual references between users within a discussion.

Civility, measured as the presence of a respectful interaction with others and the absence of insults, pejorative speech, and other markers of disrespect.

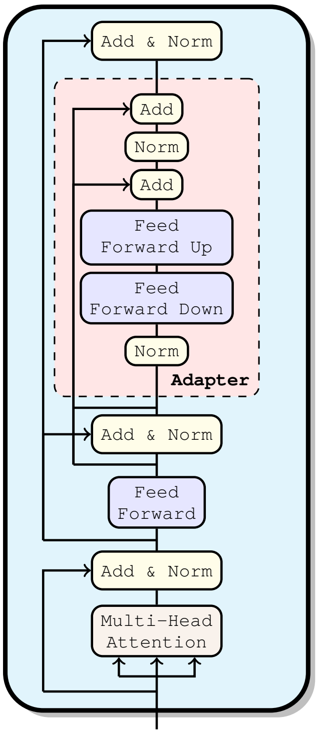

Figure 2: For the individual adapter predictions, we use a Transformer based model with adapter layers inserted after the feed forward layer of the Transformer as proposed by Pfeiffer et al. (2021).

The following coding scheme was used: all categories were coded on a four-point scale from "clearly not present" to "clearly present". Intercoder reliability was tested on a subset of 130 comments and exceeded the critical threshold of Krippendorff's α of .67 for all categories (Ø = .83). The #meinfernsehen data is annotated with the same scheme as KODIE. For #meinfernsehen intercoder reliability was tested on 159 comments, exceeding the critical threshold of Krippendorff's α of .67 for 20 out of 21 categories (Ø = .74).

We selected 19 out of the 23 deliberative quality criteria to train adapters, since some annotated aspects, e.g., *threat of violence* were not found in the data. In addition to the deliberative quality criteria, we included *storytelling*, which is considered a type II deliberation criterion, according to Bächtiger et al. (2009), since the description of personal experience when suggesting a solution contributes to the perceived quality of a comment (Falk and Lapesa, 2023b). The 20 deliberative aspects that we use are listed in Table 1. After filtering out data points with missing annotations and coding errors, we were left with a total of 13,069 comments to train our adapter models. In the following we will write

$$s_k(i) \in \{0, 1, 2, 3\} \tag{1}$$

for the k-th score $(1 \leq k \leq 20)$ of the i-th comment

	Adapter	Description	Weight
Rationality	**Relevance**	**Does the comment have a relevance for the discussed topic?**	**0.20908452**
	Fact	**Is there at least one fact claiming statement in the comment?**	**0.18285757**
	Opinion	**Is there a subjective statement made in the comment?**	**-0.11069402**
	Justification	**Is at least one statement justified in the comment?**	**0.29000763**
	Solution Proposals	**Does the comment contain a proposal how an issue could be solved?**	**0.39535126**
	Additional Knowledge	**Does the comment contain additional knowledge?**	**0.14655912**
	Question	Does the comment include a true, i.e., non-rhetoric question?	-0.07331445
Reciprocity	Referencing Users	Does the comment refer to at least one other user or to all users in the community?	-0.03768367
	Referencing Medium	Does the comment refer to the medium, the editorial team or the moderation team?	0.07019062
	Referencing Contents	Does the comment refer to content, arguments or positions in other comments?	-0.02847408
	Referencing Personal	**Does the comment refer to the person or personal characteristics of other users?**	**0.21126469**
	Referencing Format	Does the comment refer to the tone, language, spelling or other formal criteria other comments?	-0.02674237
Civility	Polite form of Address	Does the comment contain welcome or farewell phrases?	0.01482095
	Respect	Does the comment contain expressions of respect or thankfulness?	0.00732909
	Screaming	Does the comment contain clusters of punctuation or capitalization intended to imply screaming?	-0.01900971
	Vulgar	Does the comment contain language that is inappropriate for civil discourse?	-0.04995486
	Insult	Does the comment contain insults towards one or more people?	-0.05884586
	Sarcasm	**Does the comment contain biting mockery aimed at devaluing the reference object?**	**-0.15170863**
	Discrimination	Does the comment explicitly or implicitly contain unfair treatment of groups or individuals?	0.02934227
	Storytelling	**Does the commenter include personal stories or personal experiences?**	**0.10628146**

Table 1: Correlation weights w_k of all 20 trained deliberative quality adapters. The weights are calculated as the correlation coefficients between the experts' annotations and non-experts' ones. The most important indicators for a high quality comment are marked in bold. Note that positive correlations correspond to a positive trait in a high quality comment, while negative correlations correspond to negative traits.

$(1 \leq i \leq 13,069)$.

3.1.2. CrowdAnno - the Non-Experts' View

In the CrowdAnno project, Wilms et al. (2023) gathered data on non-experts' perception of uncivil, deliberative, and fact-claiming communication within German online comments through crowd annotation. The dataset includes 13,677 comments from different news media comment sections and online citizen participation projects, annotated by 681 crowdworkers. For AQuA, we used a subset of 1,742 comments that are identical to the KODIE and #meinfernsehen data. Crowd workers were tasked with evaluating, whether a comment is perceived as enriching and value-adding to the discussion or not, i.e., marking if it contains enriching communication, which could serve as a proxy for deliberative quality. The final score is aggregated from evaluations by 9 different crowd annotators via majority vote. To minimize annotator bias, the crowd workers were sampled to reflect various sociodemographic and educational backgrounds. We will write

$$c(i) \in \{0, 1\} \tag{2}$$

for the binary deliberativeness label of the i-th comment.

3.2. Training the Adapters

To automatically predict the various deliberation criteria, we use pre-trained language models, such as BERT (Devlin et al., 2019). We follow the adapter approach: adapters are extra weights θ_k, that are plugged into pre-trained language models and then learned for a specific task k. The adapted language model for the k-th deliberation criterion is written as $f_{\theta_k}(x)$, where x is some text input. Note that while learning these extra weights, we do not alter the pre-trained model weights. More precisely, we used the adapter architecture proposed by Pfeiffer et al. (2021), which is shown in Figure 2. We trained 20 individual adapters to predict scores $f_{\theta_k}(x)$ for individual indicators for deliberative quality in user comments for the KODIE dataset. For training we perform a 65% (train), 15% (val), 20% (test) split on our dataset, resulting in 8,495 training data points, 1,960 for validation and 2,614 for testing. Each of the 20 adapters for AQuA is trained with a multi-label classification objective, minimizing the cross entropy loss. We train each adapter for 10 epochs and save the model with the best macro F1 score.

3.3. Calculating the Weights

Assigning an importance to the individual quality dimensions for the overall quality measurement is not a simple task. Our intuition for weighting the deliberative criteria is to include the perception of people who potentially read and write these comments. For that reason we linked the scientific theory of deliberation to the view of non-scientists by combining the datasets described in detail in Section 3.1. More precisely, we obtain the weight for each deliberative criterion k by calculating the correlation coefficient,

$$w_k = \frac{\sum_{i=1}^{N}(s_k(i) - \bar{s}_k)(c(i) - \bar{c})}{\sqrt{\sum_{i=1}^{N}(s_k(i) - \bar{s}_k)^2}\sqrt{\sum_{i=1}^{N}(c(i) - \bar{c})^2}}, \tag{3}$$

between the scientific label $s_k(i)$ (with mean $\bar{s}_k$) for each of the $K = 20$ aspects of deliberation and the perception of crowd workers on the comments deliberativeness $c(i)$ (with mean $\bar{c}$) for all N comments. Note that w_k is a value from the interval between -1 and 1.

3.4. Building the AQuA Score

We build an overall quality score $s(x)$ for each comment as the weighted sum of the weights w_k and the predicted score $f_{\theta_k}(x)$ for each of the $K = 20$ quality adapters:

$$s(x) = \sum_{k=1}^{K} w_k f_{\theta_k}(x). \tag{4}$$

The highest and lowest possible scores depend on the number K of criteria and on the range of the predictions $f_{\theta_k}(x)$. Since the labels from KODIE are from the set $\{0, 1, 2, 3\}$, the predictions are also from this set. The highest possible score can be reached by setting all positively weighted criteria to their maximum value (i.e, 3) and all negatively weighted criteria to their minimum value (i.e, 0),

$$s_{\text{max}} = \sum_{k=0}^{K} 3 \cdot w_k \cdot [w_k \geq 0] \approx 4.9893, \tag{5}$$

where $[w_k \geq 0] = 1$ if $w_k \geq 0$ and zero otherwise. Similarly, the smallest possible score is

$$s_{\text{min}} = \sum_{k=0}^{K} 3 \cdot w_k \cdot [w_k \leq 0] \approx -1.6693. \tag{6}$$

To get a more intuitive range of values, we scale $s(x)$ to an interval between 0 and 5:

$$s_{\text{AQuA}}(x) = 5 \cdot \frac{(s(x) - s_{\text{min}})}{(s_{\text{max}} - s_{\text{min}})}, \tag{7}$$

which is the definition of our proposed AQuA score. Figure 1 graphically illustrates, how the AQuA score is calculated for a given input comment.

3.5. Applying the Score to English Comments

To apply our method to English datasets, we used the `wmt19-en-de-model`[1] (Ng et al., 2019), to automatically translate all comments in the examined dataset from English to German. Another alternative would be to train adapter models on English data. Since the KODIE dataset consists of German Facebook comments on political issues, discussing German politicians as well, we decided not to translate these comments to train adapter models, but to translate English comments and use the pre-trained German models for evaluation.

		German	Multilingual BERT	
		BERT	cased	uncased
Rationality	Relevance	**0.39**	0.37	0.37
	Fact	**0.58**	0.56	0.54
	Opinion	**0.59**	0.57	0.5
	Justification	**0.7**	0.69	0.67
	Solution Proposals	0.77	**0.79**	0.76
	Additional Knowledge	0.71	**0.78**	0.74
	Question	0.84	**0.87**	**0.87**
Reciprocity	Referencing Users	0.86	**0.88**	0.87
	Referencing Medium	0.92	0.93	**0.94**
	Referencing Contents	0.7	**0.81**	0.8
	Referencing Personal	0.83	**0.92**	**0.92**
	Referencing Format	0.89	**0.96**	**0.96**
Civility	Polite form of Address	0.96	0.97	**0.98**
	Respect	0.81	0.9	**0.91**
	Screaming	0.77	**0.81**	0.79
	Vulgar	0.76	0.74	**0.86**
	Insults	0.87	0.87	0.87
	Sarcasm	**0.48**	**0.48**	0.34
	Discrimination	0.83	**0.88**	0.87
	Storytelling	0.83	0.85	**0.86**
	Ø Total Average (F1-Score)	0.7545	**0.7815**	0.771

Table 2: Base models. We analyze the performance of different base models with adapter training on the 20 deliberative aspects. We show the weighted average F1 score. Overall, the multilingual BERT cased model performs best on the KODIE test dataset. We therefore use multilingual BERT as a base model for the AQuA score.

4. Analysis and Experiments

After defining the AQuA score in the previous sections, we briefly discuss the choice of our base model and then analyze the weights that we calculated for the individual adapter predictions. Finally, we conduct several experiments to show that our model can successfully predict deliberative quality in user comments.

4.1. Choice of the Base Model

The correlation coefficients are one important part that affect the composition of AQuA. The other part are the predictions of each of the 20 trained adapters. The adapter weights can be trained with different base architectures. To determine which base model performs best, we examine the performance of different models, namely German BERT Base cased (Chan et al., 2020) and multilingual BERT (Devlin et al., 2019) in the cased and uncased variants, on the KODIE test split. The training procedure is the same as described in Section 3.2. The results are shown in Table 2. As the datasets are highly imbalanced, and some deliberative qualities do not occur often in the training data, we report the weighted averaged F1 score, i.e., a global weighted average F1 score for each class. The trained adapter weights with the multilingual BERT model as base model outperform the German BERT model on 15 out of the 20

[1] `https://huggingface.co/facebook/wmt19-en-de`

	Label	Frequency			
		0	1	2	3
Rationality	Relevance	130	200	345	1065
	Fact	1155	113	155	317
	Opinion	27	15	13	123
	Justification	1177	78	139	346
	Solution Proposals	932	400	281	127
	Additional Knowledge	1524	76	91	48
	Question	1590	55	45	50
Reciprocity	Referencing Users	1164	128	62	386
	Referencing Medium	173	1	1	3
	Referencing Contents	1142	98	119	381
	Referencing Personal	177	1	0	0
	Referencing Format	177	0	0	1
Civility	Polite form of Address	1725	3	6	6
	Respect	1572	25	100	43
	Screaming	1612	30	53	45
	Vulgar	1654	44	23	19
	Insults	1670	29	21	20
	Sarcasm	1327	115	130	168
	Discrimination	170	2	1	5
	Storytelling	1617	59	46	18

Table 3: CrowdAnno. Absolute frequencies of each label in the subset of the CrowAnno dataset, used to calculate the correlation coefficients.

tasks. In direct comparison, the cased variant of Multilingual BERT performs slightly better than the uncased one. Based on these results we take the multilingual BERT Base cased model[2] as our base model for calculating the AQuA score.

4.2. Insights from the Correlations

The calculated correlation coefficients serve as weights in AQuA to give more importance to some deliberative aspects than others. Besides their values determining the importance for each criterion, the sign of the correlation coefficient reveals if an aspect is positively or negatively associated with comment quality. In the following, we discuss the coefficients and examine whether findings from previous deliberative research are consistent with our results. The coefficients with large absolute values are marked bold in Table 1.

For an overview of the data distribution, Table 3 lists the absolute frequencies of each label for each deliberative quality criteria in the subset of the KODIE and #meinfernsehen datasets that have been annotated using the CrowdAnno framework. These points were used to calculate the correlation coefficients. Note that these are not the frequencies in the dataset used for training the adapters. However, the small subset reflects the class imbalance that is present in the data, indicating that some categories such as vulgar language, insults and even storytelling do not occur often.

It is striking that nearly all indicators for *rationality* are strongly positively correlated with non-experts' perceived deliberative quality of comments. Using well-reasoned arguments that are relevant to the topic has been found to be an important aspect in distinguishing between comments of high and low deliberative quality (Diakopoulos, 2015; Kolhatkar et al., 2020). Unfounded expressions of opinion, on the other hand, are perceived as non-constructive, i.e., negative, in user comments. Our results support that finding, as opinion is highly negatively correlated with the perceived deliberative quality.

Of all the indicators of *reciprocity*, referring to personal characteristics of others has the greatest positive impact on the overall score. This is surprising as deliberative literature primarily highlights engaging with others' positions, not their personal traits, as a quality indicator (e.g., Ziegele et al., 2020).

Within the *civility* criteria, sarcasm stands out with a rather high negative correlation coefficient. Sarcasm, as well as doubting, criticism, and insults have been identified as one form of expressing disrespect towards other participants (Bender et al., 2011). The large correlation weight for sarcasm is a stable finding, since it is more frequent in the KODIE data, in contrast to insults.

While not being a central aspect of deliberation, storytelling in form of personal anecdotes can foster empathy and mutual understanding between participants and resolve differences (Black, 2008). Thus, it is reasonable that *storytelling* plays an important role in the weighting of AQuA, as well.

4.3. Evaluating the Score

Having trained the AQuA score using the KODIE, #meinfernsehen and CrowdAnno datasets, we next show that the learned adapter weights and correlations transfer to other datasets as well and give scores that are qualitatively and also quantitatively convincing.

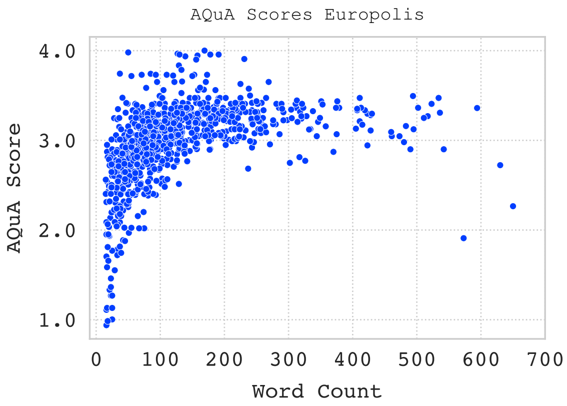

Figure 3: Europolis. AQuA scores (y-axis) vs the comment length (x-axis, word count) rule out that comment length alone is a factor for a high AQuA score.

[2] https://huggingface.co/bert-base-multilingual-cased

	Adapter	F1 Score
Rationality	Relevance	13.22
	Fact	18.48
	Opinion	42.93
	Justification	29.49
	Solution Proposals	56.04
	Additional Knowledge	38.97
	Question	62.25
Reciprocity	Referencing Users	66.85
	Referencing Medium	69.23
	Referencing Contents	66.28
	Referencing Personal	70.40
	Referencing Format	70.40
Civility	Polite form of Address	69.89
	Respect	69.67
	Screaming	67.96
	Vulgar	65.64
	Insults	70.40
	Sarcasm	66.12
	Discrimination	65.84
	Storytelling	65.33

Table 4: SOCC. Adapters that align with toxicity reach a high weighted average F1 score with toxicity levels from the SOCC dataset.

4.3.1. SFU Opinion and Comments Corpus

We predict AQuA scores on comments of the SFU opinion and comment corpus (SOCC) (Kolhatkar et al., 2020). The dataset includes 1,121 comments on news articles that have been annotated for *constructiveness* (binary annotations) and *toxicity* (four point scale from not toxic to very toxic). According to Kolhatkar et al. (2020), constructive comments are required “to create a civil dialogue through remarks that are relevant to the article and not intended to merely provoke an emotional response”.

We calculate AQuA scores and use them to predict the binary constructive label for each comment in the SOCC. Choosing a threshold of 2.3, i.e., inferring $\hat{y}_{\text{constructive}} = 1$, if $s_{\text{AQuA}} \geq$ 2.3, we get an F1 score of 81.73. Note that the threshold is a hyperparameter and a value of 2.3 was chosen, because with performed best on the data. As the dataset also comprises labels for toxic comments, we use the individual adapter predictions for *screaming*, *vulgar*, *insults*, *sarcasm*, and *discrimination* to predict the level of toxicity for each comment. Both the SOCC labels y_{toxic} as well as our predictions $s_k(i)$ are numbers from 0 to 3, therefore we simply use the individual predictions of each adapter as an indicator for the toxicity level and calculate the weighted average F1 score. With 829 comments labeled as not toxic at all (label 0), 172 with label 1, 35 with label 2 and only 7 comments that are marked as clearly toxic (label 3), the distribution is very similar to the one we see in the datasets we used for AQuA. Table 4 shows that we reach good F1 scores for adapters that align with toxicity.

4.3.2. Europolis

For a qualitative analysis of the AQuA score, we apply it to the Europolis dataset (Gerber et al., 2018). Europolis includes transcribed speech contributions of a deliberative poll on migration and climate change, annotated for *interactivity*, *respect*, *storytelling*, *justification* and *common good*. We calculate AQuA scores for each contribution in the dataset and report the top 3 highest and lowest ranked comments in Table 5. For interpretability, we list both the predicted labels of the individual adapters and the original Europolis labels (in both cases only for values greater than 0). While both differ, the AQuA labels approximately match the original Europolis labels. The top 3 comments are all rated highly with positive deliberative aspects such as storytelling, justification and additional knowledge, while the lowest comments exhibit negative deliberative aspects such as sarcasm and references to other participants. Overall, all of the the lowest scored comments are questions to clarify certain aspects in the discussion, whereas the higher scored comments consist of sophisticated opinions.

When comparing the AQuA predictions to the original Europolis labels, we find that the AQuA score seems consistent with the original labels, while enhancing the prediction since the AQuA score consists of 20 deliberative aspects instead of the 5. This demonstrates the value of AQuA as a unified score that can be applied to any dataset based on the chosen deliberative aspects.

Does comment length matter? An interesting observation is that the lowest ranked comments in the dataset are much shorter than the high ranked ones. To study whether comment length alone is the most important factor that causes our model to predict a large score, we take a closer look at the distribution of scores depending on the length of the comment. Figure 3 displays the AQuA score (y-axis) in comparison to the comment length (x-axis, word count). While it is true that short comments get the lowest scores, which is probably due to the fact that they do not have much content, the visual analysis reveals also that medium length comments get the highest scores. This rules out that comment length is the most relevant factor for our score.

5. Conclusion

In this work we introduce AQuA, an approach for an automated deliberative quality score based on large language models and adapters. The score combines annotations of experts and the view of non-experts on real online discussion comments.

We show that the trained adapters are capable

Top 3 Comments from Europolis			
Comment	**Europolis Labels**	**Adapter Predictions**	**Score**
The problem with the whole story is that first of all the cost of living has to be equalized - that includes, of course, wages, or salaries. If that - I assume we are only Poles and Germans here - and an Austrian, excuse me Julian - that we, I think, as I have come to know it - I have just said, we have a twin town in Poland - the cost of living was at least two years ago in Poland much lower than in Germany and then of course higher wages have to be paid here, so that you can buy the piece of bread, which is correspondingly lower in Poland and that's why Frankfurt/Oder to the other side is a constant border traffic. Buying gas in Poland is just much cheaper than in Frankfurt/Oder on the border. So the problem is simply that the cost of living in the individual states is so different that you can't equate it with wages and salaries at all.	*interact.: 2,* *respect: 1,* *storytelling: 1,* *justification: 2*	*rel.: 3, fact: 3,* *opinion: 3,* *justification: 3,* *suggest. sol.: 3,* *additional know.: 3,* *storytelling: 3*	4.0005
Financial problems always existed in different countries. If someone wants to live in another country, he can always do so. So if he/she wants to work a few years in some country in order to send the family money that he/she earned, he/she should not be prevented from doing so.	*interact.: 3,* *respect: 1,* *justification: 3,* *common good: 2*	*rel.: 3, fact: 3,* *justification: 3,* *suggest. sol.: 3,* *additional know.: 2,* *storytelling: 1*	3.9803
Many people are coming to other countries not just because of economic reasons. Often, they are persecuted in their own countries on the religious grounds and they are trying to find asylum in another country. Then, the government should give them political asylum, papers or right of permanent residency and then they can work. For example Germany is rich enough to give jobs for immigrants and integrate them in the society because the society is aging and somebody has to work for the new generation which would like to get future pensions or something like that. Society is aging so they need immigrants. Similar to Poland where the government should legalize immigrants in a similar way. It is hard to say how it actually should look like.	*respect: 2,* *justification: 2,* *common good: 1*	*rel.: 3, fact: 3,* *suggest. sol.: 3,* *additional know.: 2.,* *justification: 3,* *discrim.: 3*	3.9666

Lowest 3 Comments from Europolis			
Comment	**Europolis Labels**	**Adapter Predictions**	**Score**
A question for Udo: To what dimension is the problem with the migration of workers growing?	*interact.: 2,* *respect: 1*	*question: 3,* *ref. user: 3,* *ref. content: 3*	0.9393
Thank you very much. Aurore, you also wanted to say something especially before the break but now too?	*interact.: 2,* *respect: 1,* *storytelling: 1,* *justification: 2,* *common good: 2*	*fact: 1, question: 3,* *ref. user: 3,* *ref. content: 3,* *polite addr.: 2,* *sarcasm: 1*	0.9849
To tell you the truth, I do not know what is discussed? Are we talking about the quotas – how many people could come here?	*respect: 1,* *storytelling: 1,* *justification: 1,* *common good: 1*	*question: 3,* *ref. user: 3*	1.0034

Table 5: Europolis. Top 3 comments with the highest and top 3 comments with the lowest calculated AQuA scores. We only show the scores and the predicted labels of the individual adapters where the prediction is larger than zero. The original labels (from Europolis, 5 labels) show that the AQuA score is well aligned with the original labels.

of predicting individual scores for different aspects of deliberative quality and that the overall score aggregates these predictions in a meaningful way. The correlation coefficients between experts' and non-experts' annotations reveal the most important positive and negative deliberative aspects, which allows us to confirm theoretical and empirical findings in deliberation literature into AQuA.

Furthermore, we evaluate our score (trained on KODIE and CrowdAnno) on two further datasets (SOCC and Europolis) to show that the predictions of the learned adapters transfer well to unseen datasets. First, we show that the adapter predictions that build the AQuA score are useful for classifying constructive and toxic comments on the SOCC dataset. Then we perform a qualitative analysis of the AQuA score by manual assessing the top 3 and bottom 3 scored comments in the Europolis dataset and show that comments with well formed opinions receive large scores, while comments providing little value to the discussion receive lower scores.

Overall, we show that AQuA can be used successfully to automatically assess deliberative quality while aligning with theoretical and empirical background in deliberation literature.

6. Bibliographical References

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189.

André Bächtiger, Simon Niemeyer, Michael Neblo, Marco R. Steenbergen, and Jürg Steiner. 2009. Disentangling diversity in deliberative democracy: Competing theories, their blind spots and complementarities. *Journal of Political Philosophy*, 18(1):32–63.

André Bächtiger, Susumu Shikano, Seraina Pedrini, and Mirjam Ryser. 2009. Measuring deliberation 2.0: standards, discourse types, and sequenzialization. In *ECPR General Conference*, pages 5–12. Potsdam.

Nick Beauchamp. 2020. 321Modeling and Measuring Deliberation Online. In *The Oxford Handbook of Networked Communication*. Oxford University Press.

Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. 2011. Annotating social acts: authority claims and alignment moves in wikipedia talk pages. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, page 48–57, USA. Association for Computational Linguistics.

Laura W Black. 2008. Listening to the city: Difference, identity, and storytelling in online deliberative groups. *Journal of Deliberative Democracy*, 5(1).

André Bächtiger, Marlène Gerber, and Eléonore Fournier-Tombs. 2022. 83Discourse Quality Index. In *Research Methods in Deliberative Democracy*. Oxford University Press.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Gina Masullo Chen. 2017. *Online incivility and public debate: Nasty talk*. Springer.

Kevin Coe, Kate Kenski, and Stephen A. Rains. 2014. Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments. *Journal of Communication*, 64(4):658–679.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicholas Diakopoulos. 2015. Picking the nyt picks: Editorial criteria and automation in the curation of online news comments. *ISOJ Journal*, 5(1):147–166.

John S Dryzek. 2002. *Deliberative democracy and beyond: Liberals, critics, contestations*. Oxford University Press, USA.

Katharina Esau, Dannica Fleuß, and Sarah-Michelle Nienhaus. 2021. Different arenas, different deliberative quality? using a systemic framework to evaluate online deliberation on immigration policy in germany. *Policy & Internet*, 13(1):86–112.

Neele Falk and Gabriella Lapesa. 2023a. Bridging argument quality and deliberative quality annotations with adapters. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2469–2488, Dubrovnik, Croatia. Association for Computational Linguistics.

Neele Falk and Gabriella Lapesa. 2023b. StoryARG: a corpus of narratives and personal experiences in argumentative texts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2350–2372, Toronto, Canada. Association for Computational Linguistics.

Eleonore Fournier-Tombs and Giovanna Di Marzo Serugendo. 2020. DelibAnalysis: Understanding the quality of online political discourse with machine learning. *Journal of Information Science*, 46(6):810–822.

Dennis Friess and Christiane Eilders. 2015. A systematic review of online deliberation research. *Policy & Internet*, 7(3):319–339.

Dennis Friess, Marc Ziegele, and Dominique Heinbach. 2021. Collective civic moderation for deliberation? exploring the links between citizens' organized engagement in comment sections and the deliberative quality of online discussions. *Political Communication*, 38(5):624–646.

Marlène Gerber, André Bächtiger, Susumu Shikano, Simon Reber, and Samuel Rohr. 2018. Deliberative abilities and influence in a transnational deliberative poll (europolis). *British Journal of Political Science*, 48(4):1093–1118.

Frauke Gerlach and Christiane Eilders, editors. 2022. *#meinfernsehen 2021*. Nomos, Baden-Baden.

Valentin Gold, Mennatallah El-Assady, Annette Hautli-Janisz, Tina Bögel, Christian Rohrdantz, Miriam Butt, Katharina Holzinger, and Daniel Keim. 2015. Visual linguistic analysis of political discussions: Measuring deliberative quality. *Digital Scholarship in the Humanities*, 32(1):141–158.

Todd Graham. 2010. The use of expressives in online political talk: Impeding or facilitating the normative goals of deliberation? In *Electronic Participation*, pages 26–41, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dominique Heinbach, Lena Wilms, and Marc Ziegele. 2022. Effects of empowerment moderation in online discussions: A field experiment with four news outlets. In *72nd Annual Conference of the International Communication Association (ICA)*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. 2020. The sfu opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, 4:155–190.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

John Mendonca, Alon Lavie, and Isabel Trancoso. 2022. QualityAdapt: an automatic dialogue quality estimation framework. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 83–90, Edinburgh, UK. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Simon Niemeyer, Francesco Veri, John S. Dryzek, and André Bächtier. 2024. How deliberation happens: Enabling deliberative reason. *American Political Science Review*, 118(1):345–362.

Zizi Papacharissi. 2004. Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, 6(2):259–283.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. 2023. Adapters: A unified library for parameter-efficient and modular transfer learning.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Mary F Scudder. 2022. Measuring democratic listening: A listening quality index. *Political research quarterly*, 75(1):175–187.

Bokyong Shin and Mikko Rask. 2021. Assessment of online deliberative quality: New indicators using network analysis and time-series analysis. *Sustainability*, 13(3).

Marco R Steenbergen, André Bächtiger, Markus Spörndli, and Jürg Steiner. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1:21–48.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lena Wilms, Anke Stoll, Marc Ziegele, and Katharina. Gerl. 2023. Bildungsbezogene Biases in crowd-annotierten Daten zur automatischen Klassifikation von konstruktiven und inzivilen Kommentaren (Educational biases in crowd-annotated data for the automatic classification of constructive and incivil comments). In *Annual Conference of the Political Communication Devision of the German Association of Communication Science (DGPuK)*.

Dominik Wyss, Simon Beste, and André Bächtiger. 2015. A decline in the quality of debate? the evolution of cognitive complexity in swiss parliamentary debates on immigration (1968–2014). *Swiss Political Science Review*, 21(4):636–653.

Marc Ziegele, Oliver Quiring, Katharina Esau, and Dennis Friess. 2020. Linking news value theory with online deliberation: How news factors and illustration factors in news articles affect the deliberative quality of user discussions in sns' comment sections. *Communication Research*, 47(6):860–890.