Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3301–3318
Marseille, 20-25 June 2022

Scaling up Discourse Quality Annotation for Political Science

Neele Falk, Gabriella Lapesa
Institute for Natural Language Processing, University of Stuttgart
Pfaffenwaldring 5b 70569 Stuttgart
name.surname@ims.uni-stuttgart.de

Abstract

The empirical quantification of the quality of a contribution to a political discussion is at the heart of deliberative theory, the subdiscipline of political science which investigates decision-making in deliberative democracy. Existing annotation on deliberative quality is time-consuming and carried out by experts, typically resulting in small datasets which also suffer from strong class imbalance. Scaling up such annotations with automatic tools is desirable, but very challenging. We take up this challenge and explore different strategies to improve the prediction of deliberative quality dimensions (justification, common good, interactivity, respect) in a standard dataset. Our results show that simple data augmentation techniques successfully alleviate data imbalance. Classifiers based on linguistic features (textual complexity and sentiment/polarity) and classifiers integrating argument quality annotations (from the argument mining community in NLP) were consistently outperformed by transformer-based models, with or without data augmentation.

Keywords: annotation, evaluation, machine learning, argumentation, political science

1. Introduction

How can we arrive at better political decisions? This question is at the heart of deliberative theory (Fishkin, 1995), the subfield of the political sciences which investigates decision-making in deliberative democracy (e.g., in parliamentary discussions). The crucial focus of deliberative theory is on the role of the discourse exchange that leads up to the actual decision: deliberation is not only the output of the decision making, but also the discussion that leads up to it. The basic assumption of deliberative democracy is that a rational, respectful exchange of different perspectives and arguments on an issue can lead to better decisions. This leads to the empirical question of measuring discourse quality. Not only is the empirical quantification of discourse quality relevant to get a better theoretical understanding of the dynamics of democratic decision making, its real-life application potential is invaluable when it comes to supporting deliberation processes which take place not only in parliaments, but also in groups of lay people discussing societally relevant problems (Vecchi et al., 2021).

Unsurprisingly, research in deliberative theory has established the conceptual coordinates and annotation schemas to investigate discourse quality empirically, leading to the formulation of the notion of Deliberative Quality. Different sub-dimensions of deliberative quality capture, for example, whether the participants in the discourse treat each other with respect, take up other opinions and sufficiently justify their positions. The best-known framework on the basis of which many discourse data is annotated is the discourse quality index (DQI) (Steenbergen et al., 2003), which defines the various aspects of discourse quality. Annotation is time-consuming and requires highly trained annotators: it is thus no wonder that the available datasets are small. Besides, the employed annotation schemas tend to be fine-grained with multiple values for each sub-dimension and an often highly imbalanced distribution of labels.

Given the core role played by the quantification of discourse quality in political science, it goes without saying that automating its annotation with Natural Language Processing tools has the potential of scaling up the scope of the analysis of deliberative phenomena. This paper investigates the extent to which automatic modeling of discourse quality based on the DQI is possible. We conduct our experiments on the reference dataset for deliberative theory, *Europolis* (Gerber et al., 2018); we evaluate different model architectures (feature-based vs. transformer models) and explore alternative strategies to cope with the low-resource and class imbalance issues.

The overarching goal of our experiments is to *integrate the existing annotation* to improve the automatic prediction. In the first set of experiments, presented in section 4, we support automatic prediction in two alternative ways: a) we annotate the dataset with linguistic properties (textual complexity and sentiment/polarity) using off-the-shelf tools and use them as features for an automatic classifier; b) we use data augmentation, a technique commonly employed in Machine Learning to achieve a better class distribution during training, and we test its impact on the performance of a standard transformer text-based model. We show that text-based models outperform the feature-based ones and crucially, that data augmentation, although potentially noisy, further improves the results on two DQ dimensions out of four.

In the second set of experiments, reported in section 5, we explore an alternative strategy to support automatic DQ prediction, namely the integration of annotations of

a compatible, yet not overlapping, phenomenon from another discipline: Argument Quality (AQ), as conceptualized and annotated in computational argumentation research in NLP (Wachsmuth et al., 2017a; Wachsmuth et al., 2017b; Wachsmuth and Werner, 2020; Toledo et al., 2019; Gretz et al., 2020). Comparably to DQ, several aspects are considered when annotating Argument Quality. For example, is the argument clearly stated? Are there sufficient justifications? Is the content relevant to the topic? Larger amounts of data (some human-, some automatically-annotated) are available, as well as automatic tools. An open question is to what extent DQ and AQ can be integrated. If there is an overlap, can existing data and models for capturing AQ lead to improved modeling of DQ? We perform an annotation study of AQ on the Europolis dataset and conduct modeling experiments to test whether the prediction of DQ (feature-based, text-based with or without augmentation) can be improved by incorporating AQ. We show that at the human annotation level there is an overlap between AQ and DQ, as AQ correlates with the rationality and respect subdimensions of DQ. At the modeling level, integrating AQ does not improve the prediction of DQ in the best-performing model (text-based, augmented) but it does improve the feature-based model.

The contributions of our work are at different levels: At the level of **experimental results**, we show how simple data augmentation techniques can be used to alleviate class imbalance, an issue that high-quality annotation which has not been carried out with the purpose of computational modeling always faces; to the best of our knowledge, we are the first one to apply this solution to DQ data. At the **methodological level**, we employ model introspection techniques to get a better understanding of what drives the performance of our classifiers. At the level of the **theoretical insights**, we make the first steps into the comparison and integration of AQ and DQ – with interesting insights at the annotation level, and mixed results at the modeling level.[1]

2. Related Work

Deliberative Quality The Discourse Quality Index (DQI) has been introduced by Steenbergen et al. (2003) to enable a quantitative analysis of discourse quality and has since provided the basis for an empirical investigation of deliberative theory, especially as it has been supported by the deliberative theorists (Habermas, 2005; Thompson, 2008). The index defines various standards for good deliberation, focusing on a rational, respectful exchange of relevant arguments. The automatic prediction of specific dimensions of the DQI has been tackled in Fournier-Tombs and Di Marzo Serugendo (2019) and Fournier-Tombs and MacKenzie (2021). These works introduce a framework called *DelibAnalysis*, which uses a feature-based approach to predict different dimensions of the DQI. The tool is evaluated on different datasets (from legislative debates to social media). However the classification task is simplified, as each dimension is converted into a binary classification problem (a certain quality dimension is 'activated' or not), opposed to this work, in which we try to model the more fine-grained distinctions for each quality dimension.

Argument Quality In Argument Mining, recent work has lead a research direction that deals with a theory-based definition and automatic modeling of argument quality. Wachsmuth et al. (2017b) have developed a taxonomy of different aspects of argument quality that is informed by existing theories. This taxonomy breaks down argument quality into three core dimensions: COGENCY reflects the logical coherence of an argument and whether there are sufficient reasons to draw the conclusion; EFFECTIVENESS refers to rhetorical aspects that make an argument more or less persuasive, for example, credibility of the author or appropriate use of language and emotion; REASONABLENESS considers the argument in a global context and evaluates whether it contributes to the issue's resolution and is acceptable by the other discourse participants. This new definition has been the basis for the creation of new corpora from different domains (Ng et al., 2020), where feature-based (Wachsmuth and Werner, 2020) and neural models were tested for automatic prediction (Lauscher et al., 2020).

In contrast to AQ, DQ puts the emphasis on aspects that favor an equality-oriented and appreciative discourse. There is, however, an overlap between AQ and the DQI, as the DQI also contains a dimension that measures rational argumentation. Our work aims to shed light on the potential overlap by annotating an existing corpus of DQ with AQ scores in order to find out the exact dimensions of AQ and DQ that correlate.

3. Data

In this work we use *EuroPolis* (Gerber et al., 2018), a dataset which was created by deliberative theorists to investigate different dimensions of deliberative quality empirically. The data consists of transcribed speeches of small group-discussions that were part of a transnational poll which took place in Brussels in 2009. The group-discussions were translated simultaneously, as within each group between 2-5 different nationalities participated. A sample of transcriptions of 13 groups was annotated by political scientists according to an updated version of the Discourse Quality Index (DQI). The language of the transcriptions was either German, French or English and the topic under discussion was immigration. After filtering out all contributions with less than 30 tokens, the length of the contributions ranges between 30 and 1000 with a mean of 157 tokens.

We focus on the following four dimensions of deliberative quality; each dimension consists of several levels

[1]The data(splits) and the code are available at `https://github.com/Blubberli/empiricalDQI.git`.

that can be arranged more or less precisely on a continuum corresponding from low to high quality.

Justification: This dimension encodes whether the speaker provides a complete justification and to what extent (do they only illustrate the problem or provide an in-depth reflection and several reasons to justify?).

Common good: This dimension captures whether or not an argument is framed with regards to the 'common good'. Common good is identified narrowly with the home country of the speaker or with the interests of a broader community (e.g. European interests) or of a more abstract collective (in the spirit of solidarity or equality).

Interactivity: This dimension quantifies the extent to which the speaker considers (i.e., makes reference to) other participants' contributions arguments and whether they value or disparage these.

Respect: This dimension measures whether the speaker shows empathy towards other groups (e.g. immigrants), whether they show explicit respect towards them or degrade them.

In order to use a monolingual model for predicting the deliberative quality of the speeches, we translated the German and French contributions into English, using DeepL.[2] The quality of the translation has been checked by a native speaker each for French and German in order to make sure that the translation conveyed the original message and that the output was grammatical.

To carry out our experiments, we slightly simplified the original annotation schema: very low-frequency labels on some dimensions were either discarded or merged with another class if possible. Table 1 shows an overview of the labels for each dimension with the absolute frequency in the dataset; the labels are ordered from the level expressing the lowest to the level corresponding to the highest quality. Especially COMMON GOOD and RESPECT suffer from severe class imbalance. Excerpts from the Annotations Guidelines of DQ can be found in the appendix, section 8.1

Table 2 shows two examples from the Europolis dataset with different levels for each quality dimension. The examples also illustrate that a high level on one dimension does not necessarily correspond to a high level on another. The first example has a high level of JUSTIFICATION (*qualified just.*) and a *reference to common good* but but does not address the contributions of other participants (*no reference*); the second example on the other hand contains a *positive reference* to another discourse participant but is at the same time *disrespectful* towards certain groups (immigrants).

justification		**common good**	
no just.	138	no reference	128
inferior just.	372	own country	675
qualified just.	303	reference to c.good	107
sophisticated just.	97		
total	**910**	**total**	**910**
interactivity		**respect**	
negative reference	40	disrespectful	79
no reference	380	implicit	657
neutral reference	324	explicit	126
positive reference	166		
total	**910**	**total**	**862**

Table 1: Overview of the dimensions of Deliberative Quality in *Europolis* with frequency for each class for each dimension. The last row shows the total number of instances for the corresponding dimension.

4. Exp 1: Linguistic features & Data augmentation

Our task is to predict the deliberative quality of a specific dimension given the (transcribed/translated) spoken contribution as input.

We pre-process the transcription by stripping off the time stamps and removing empty lines, tabs and links. We carry out stratified five-fold cross-validation, so every data point is tested and the class distribution for training and test set is similar. For each split, 60% of the data is used for training and 20% as validation and test set. We report the results on each of the four dimensions for the following models (implementation details and hyperparameter settings of the models can be found in the appendix, section 8.3) :

1. **Baseline (majority)** This baseline predicts the majority class. The results show which of the dimensions suffer most from class imbalance.
2. **Feature-based (feats tree)** A tree-based ensemble using gradient boosted trees with linguistic features such as textual complexity and sentiment described in section 4.1
3. **Text-based (roberta-base)**: We fine-tune `roberta-base` with a multi-class classification head on top. We fine-tune all parameters and train the model for a maximum of 15 epochs. We use early-stopping and pick the model that achieves the highest F1-macro score on the validation set.
4. **Text-based, augmented (roberta-augment)**: a version of `roberta-base` employing data-augmentation during training, as discussed in section 4.2.

4.1. Feature-based classifier: linguistic features

We automatically enrich the existing dataset with features of textual complexity and sentiment which can serve (a) as a basis for quantitative analyses of relationships between DQ and these linguistic properties and (b) as additional input to machine learning models. We extract the following types of features:

Textual complexity (19): features of lexical diversity (e.g. type-token ratio), and lexical sophistication,

[2] (`https://www.deepl.com/translator`)

contribution	just	c.good	int	resp
In fact, that's a big problem. The countries of the EU need to cut military spending and put the money aside for the welfare state, for pensions, and for European civilization, European civilization, which is very human and humane. We don't need military, we don't need weapons, that's one way to address the problem.	*qualified*	*common good*	*no reference*	*implicit*
I would like to add something that I think is very important: It's good that the immigration of the 20th and 21st century is the immigration of people from cultures completely opposite to ours. And this is the problem. These people don't care about our culture at all. And that's the problem that we try never to develop, never to raise them, we try to say what Madame raised earlier. We had problems, these people are different.	*inferior*	*no reference*	*positive reference*	*disrespectful*

Table 2: Contributions from *Europolis* with annotation for the deliberative quality dimensions JUSTIFICATION (just), COMMON GOOD (c.good), INTERACTIVITY (int) and RESPECT (resp).

such as ngram-frequencies or contextual distinctiveness. Deeper levels of justification can be expected to correlate with more complex language use.

Sentiment/Polarity (20): we extracted different features related to polarity and sentiment using a lexicon-based approach. These features do not only include information about polarity and emotions, but also about social order, cultural values and beliefs which are especially useful for the analysis of deliberative discussions. For more details about each feature refer to appendix section 8.2.

Table 3 compares the original training size and that of the augmented data for each dimension (average over all splits). The classes in the final training data are balanced, however note that the size of the training data now differs, as it depends on the training frequency of the minority class.

4.2. Text-based classifier: data augmentation

Because of the size and the class imbalance of this dataset, we try a simple data augmentation method. We enrich the training data for each quality dimension with Easy Data Augmentation (EDA) (Wei and Zou, 2019). This method creates synthetic data applying four different heuristics: synonym replacement (replace words of a sentence with their synonyms), random insertion (insert a random synonym of any word of the sentence into a random position), random swap (randomly swap two words in a sentence) and random deletion (randomly remove a word of a sentence). Creating synthetic training data with the help of random perturbation operations has been effectively applied in other works to improve performance in the low-resource scenario, e.g. to social media (Ansari et al., 2021) or for named-entity recognition (Dai and Adel, 2020) but to the best of our knowledge has not yet been applied to data from political science.

For each comment in the dataset we create 10 augmented comments applying all four heuristics[3]. Finally for each training set we create an augmented dataset by drawing comments from the augmented database randomly for each quality dimension until all classes are balanced. Then we add as many additional examples for each class as available (as many comments available for the lowest-frequency class). Consider for instance a training set for INTERACTIVITY of one split with the following class distribution: *negative reference*: 25, *no reference*: 234, *neutral reference*: 191, *positive reference*: 96. As a first step the the dataset is augmented until all classes have at least 234 instances (corresponding to frequency of the majority class). The minority class is *negative reference* with 25 instances, thus for this class only a total of 250 augmented instances is available. After 209 instances have been added to the dataset in the first step the remaining possible number of instances per class that can be added to the augmented dataset is 41.

[3]We use `https://github.com/jasonwei20/eda_nlp` and apply the heuristics to 10% of each comment.

	orig. size	augm. size	classes
justification	546	2,523	4
common good	546	2,031	3
interactivity	546	1,020	4
respect	522	1,663	3

Table 3: Training size (original vs. augmented) and number of classes (*classes*) for each deliberative dimension

	just	c.good	int	resp
majority	0.15	0.28	0.15	0.29
feats tree	0.36	0.41	0.27	0.40
roberta-base	**0.74**	**0.86**	0.65	0.75
roberta-augment	0.71	0.81	**0.80**	**0.84**

Table 4: F1 macro score for the models trained and tested on each DQ: just[ification], c[ommon].good, int[eractivity], resp[ect].

4.3. Results

Table 4 shows the F1-macro score for each dimension. The feature-based classifier improves over the majority class, the greatest improvement can be observed for JUSTIFICATION. The transformer-based models achieve good results for all dimensions, and this is reassuring given that the dataset is quite small and the classes are imbalanced. Data augmentation further improves classification performance for the most imbalanced dimensions, RESPECT and INTERACTIVITY, with an increase of 9 and 15% F1. On the other hand, the performance slightly drops for JUSTIFICATION and COMMON GOOD when trained on the augmented data although they have the largest augmented training set.

One possible explanation is that the ratio between high quality training instances and noise is no longer balanced enough. Especially for JUSTIFICATION, it is likely that the perturbations have affected the quality of the representation. One possible future step would be to try meta-learning or instance-weighting strategies to sample the augmented instances more effectively (Yi et al., 2021; Mou et al., 2021).

In what follows, we further investigate the factors which drive the performance of our classifiers. Even if the transformer-based classifier clearly outperformed the feature-based one, we decided to further investigate the performance of both as they provide different perspectives on our dataset.

The most salient features of the feature-based model can reveal the most salient linguistic properties of the individual DQ dimensions and thus can give an empirical picture of the underlying dataset. For that purpose we extract feature attributions to analyze the impact of the different properties on the model predictions. We use SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) to extract features of a high impact for the different DQ dimensions for the tree-based model. [4] SHAP is based on the game-theoretically optimal Shapley values. The SHAP values show the average marginal contribution of a feature across all possible coalitions. This allows for a visualization like the one in Figure 1, where the most influencing features for each dimension are represented as a stacked bar plot, allowing feature importance to be distinguished for the individual classes.

On the other hand, the most salient features of the transformer model can be interpreted as the most salient lexical properties of the data. For the transformer-based models we use the library `transformer-interpret`, which uses 'Layer Integrated Gradients (LIG)'(Sundararajan et al., 2017) to compute importance scores for the input words. We compute attributions for all words in all test sets setting a threshold of 0.1 to extract the words having a positive impact on the class probability. For a global picture we then look at the most frequent words that have high attribution scores (c.f. Tables 17,18, 19, 20 in section 8.4 of the appendix, for a list of words frequently retrieving a high attribution for each dimension).

4.4. Analysis

In what follows, we present the general trends for each DQ dimension, first for the feature-based model, and then for the transformer model.

Justification The analysis of the loading of different features for JUSTIFICATION in the feature-based model, illustrated in Figure 1, reveals a length bias towards the level of justification, as the number of words is the most important feature that is picked up by the model for the extreme classes: a very low or very high level of justification. Looking at the more detailed beeswarm plots for each class, displayed in the appendix in section 8.4 [5], reveals that we obtain a lower probability for *sophisticated justification* (c.f. Figure 4d) when having shorter comments. Shorter contributions are more likely to get a high probability for no justification (c.f. Figure 4a). Higher values of lexical diversity correlate with a higher probability for qualified and *sophisticated justification* and a lower probability for *no justification* indicating that the model picks up on the diversity of the vocabulary when predicting the level of justification.

Let us now turn to the the most salient words picked up by the transformer model. We can observe that the models learn to associate questions with *no justification* ('question', 'what', '?') while the use of causal connectives and words expressing policy actions give a higher probability to the classes of higher justification ('because', 'should', 'must') (c.f. Table 17). On top of that, political words are more often associated with a higher level of justification ('population', 'migration', 'national', governments'). For instance we can see that in example 2a connectives like 'because' and 'finally' as well as expressions of opinion ('I am in favor', 'These are two reasons') have a positive impact. In this example we can also see that specific words associated to the topic the argument is about obtained high attribution scores ('migration', 'work', 'jobs', 'employers').

Common good. For COMMON GOOD, we can observe that the features expressing a specific domain (e.g. economy / social order words) are the most important. The beeswarm plots (Figure 5) show that a high amount of politeness words and words expressing an ethical or political concept ('social order') increase the probability of *reference to common good*. High amounts of politeness words on the other hand predict a lower probability for reference to *own country* (which is at a lower range of deliberative quality with respect to reference to common good). The model therefore learns to associate a higher level of politeness to a higher deliberative quality of COMMON GOOD.

Looking at the most salient lexical cues picked up by the transformer-based model (c.f. Table 18), we also find that political words and words regarding the more general collective get high attributions, for example 'Europe', 'Europeans', 'global', 'member', 'we'. Comments containing common good references regarding their *own country* often talk about (il)legality, work, integration and culture ('people', 'legal', 'work', 'language', 'country'), probably pointing out problems or positive examples of immigration based on experiences in their own country. The most relevant words for *no reference to common good* (which corresponds to a lower quality) are mentions of specific countries (Germany, France, Poland) but also similar words as for

[4] We use the python package `shap` and the TreeExplainer to compute the SHAP values.

[5] Figure 4,5,6,7 illustrate the positive and negative relationships and importance of a feature towards each class for each dimension.

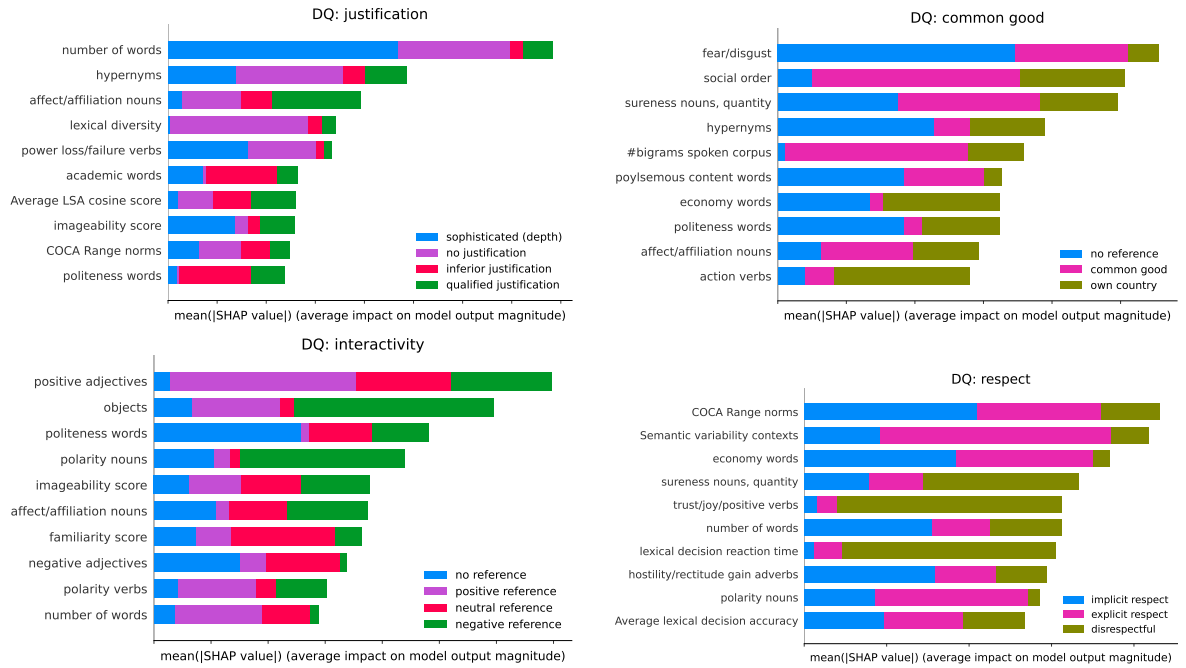


Figure 1: SHAP feature importance

#s Strange ly enough , I am in favor of strict border control , because illegal migration compromises the position of migrants who are already in Europe . So you have to see that this illegal migration , the reason for illegal migration , has to be eliminated . These are two reasons why people come . Because there is an opportunity for work and because there are too few people in Europe to do certain jobs . And so if you are in the EU and you want to stay there , you have to see that employers don 't have that kind of need so that the market for illegal immigrants is finally dried up . #/s

(a) JUSTIFICATION: example for *sophisticated justification*

#s Well , I agree with all those who had a say , but what is important for me is that also the people who are here illegally and who have very often , for example , health problems and no money to go to see a doctor - This is really nothing good - neither for them nor for their families and it is a reason for suffering . #/s

(b) INTERACTIVITY: example for *positive reference to other participants*

Figure 2: Examples of Europolis with highlighted important words picked up by the best model: *green* words have a positive effect on the corresponding class and *red* words a negative effect.

own country ('problem(s)', 'people'). Interestingly the most relevant function words contain 'their' and 'our' indicating that these comments tend to see immigration as a problem and contrast themselves and the 'others' (the immigrants). The example 8a in the appendix highlights the main point of the contribution representing an ethical aspect ('EU countries should contribute a part of the generated wealth and welfare to the poorer') but also sub-clauses that structure the argument ('as far as the discussion is concerned').

Interactivity The most important features for INTERACTIVITY are positive and negative adjectives which have the expected effects on the corresponding classes expressing the polarity (high amount of positive adjectives predicts a high probability for a *positive reference* to other participants). A high amount of concrete words (captured by imageability) predicts a higher probability for *negative reference* (Figure 6b). The reasons for this correlation require further investigation of the vocabulary's concreteness scores of the arguments associated with this level of interactivity.

Looking at the transformer-based models reveals a similar pattern as for the feature-based models: they are picking up positive and negative words for the extreme classes (positive vs. negative references to other participants), for example 'agree', 'good', 'yes', 'right' for positive references and 'disagree', 'not', 'nothing' for negative references (Table 19), which also shows that the models pick up on affirmatives and negatives. This is also illustrated in example 2b which retrieves a strong attribution for 'well I agree'.

Respect Figure 1 shows that linguistic complexity is predictive of comments that are *disrespectful* or *implicitly* respectful. The beeswarm plot (c.f. Figure 7a) shows that disrespectful comments have a rather low complexity (low type token ratio predicts a higher

probability for this class) and are less lexically sophisticated[6] showing that the speakers use more general and frequent words when being disrespectful. For explicit respect we find a higher amount of positive adjectives and a higher lexical diversity to be predictive (c.f. Figure 7c).

The transformer-based models picks up the words (c.f. Table 20) associated with the expressed stereotypes in the case of disrespectful comments ('immigrants', 'problem', 'integrate', 'here'), as the corresponding speakers talk about the immigrants as a threat or a problem. However disrespectful comments seem to be quite implicit as we cannot observe a pattern of high attributions for clearly negative adjectives or nouns. For comments of explicit respect the models pick up on words that offer explanations or reasons for why people immigrate ('live', 'problems', 'better', 'money') and words regarding the social good ('Europe', 'human'). Looking at concrete examples reveals that indeed for comments with a high probability for *disrespect* there are no clear attributions but negative and positive attributions are distributed across the whole comment. The example 8b in the appendix shows that 'image of women' contributes positively to the prediction of *disrespect* but 'clearly different' pushes the probability for that class in the other direction. It would therefore be necessary to have a more detailed look at explanations for the examples of this dimension to uncover potential biases picked up by the model (e.g. do words like 'Islam' or 'immigrants' bias the model towards the prediction of a certain class?).

5. Exp. 2: Integration of AQ and DQ

The different frameworks from AQ and DQ offer alternative possibilities to measure the quality of argumentation and discourse. With an annotation study we aim to uncover the areas of overlap between the theories. Since, unlike for the measurement of DQ, there exist already larger data and models to automatically measure AQ, this information could be used to support the automatic assessment of DQ and serve as additional help for the low-resource problem.

5.1. Enriching Europolis with Argument Quality: Human Annotation

For our pilot study we select a subset of Europolis and collect human annotations for AQ using the guidelines developed in Ng et al. (2020). The guidelines are based on a simplified taxonomy of AQ (Wachsmuth et al., 2017a) and define the core dimensions, COGENCY, EFFECTIVENESS, REASONABLENESS as well as an aggregated score for OVERALL quality[7]. We gather AQ annotations for 105 comments by two annotators, both students of computational linguistics, trained on annotating AQ in other studies. The first annotator is an English native speaker, the second annotator is German with a high proficiency in English. The comments were annotated based on English, the original language of the transcriptions. As a first step, as it is common in Argument Mining, the annotators identified whether a contribution is argumentative or not. Table 5 shows the Pearson correlation and the weighted Cohen's kappa between the two annotators for each AQ dimension and the OVERALL AQ score for the argumentative contributions. We observe a low agreement for all dimensions with COGENCY and EFFECTIVENESS being the most difficult.

cog	eff	reas	overall	measure
0.31	0.27	0.40	0.41	r
0.14	0.15	0.18	0.20	κ

Table 5: Agreement (Pearson correlation (r) and weighted Cohen's kappa (κ)) for each AQ dimension (cog[ency], eff[ectiveness], reas[onableness]) on the two-annotator subset (84 comments) of *EuropolisAQ*.

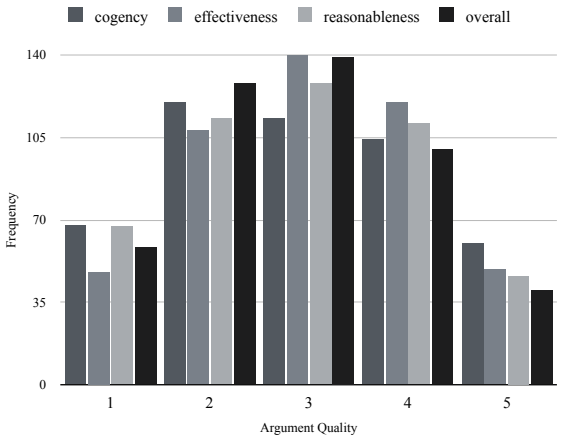


Figure 3: Histogram of the AQ scores for *EuropolisAQ*

We proceeded to annotate with the German annotator all other Europolis contributions whose transcriptions were in English or German, in their underlying original language (either English or German). The final dataset annotated with AQ consists of 534 contributions that were first annotated with *argumentative* vs. *non-argumentative* label. The argumentative subset, further referred to as *EuropolisAQ* contains 513 comments with scores for both DQ (original annotation) and AQ (our annotation). Figure 3 shows the distribution of the AQ scores for *EuropolisAQ*. As expected extreme cases of very high or very low quality (1 or 5) are less frequent and most of the contributions were rated between 2 and 4 for all dimensions. An interesting observation is that COGENCY is more likely to be rated with the minimum or maximum value.

Table 6 shows the Pearson correlation between each

[6]High LSA cosine score and low number of hypernyms predict a high probability for this class, so the vocabulary used in these comments is more general and occurs in a wide range of different contexts.

[7]The concrete definitions for these dimensions can be found in Table 12, section 8.1 in the appendix

	just	c.good	int	resp
cog	0.33^*	0.12	-0.01	0.34^*
eff.	0.26^*	0.07	0.10	0.25^*
reas.	0.27^*	0.12	0.12	0.23^*
overall	0.30^*	0.11	0.06	0.30^*

Table 6: *EuropolisAQ*: Correlation between AQ (cog[ency], eff[ectiveness], reas[onableness]) and DQ (just[ification], c[ommon].good, int[eractivity], resp[ect])

AQ and DQ dimension[8]. We can observe significant positive correlations between justification and all AQ dimensions. We find the highest correlation to cogency, which makes sense as the logical dimension of AQ and the rational level of DQ were expected to have the highest overlap. We also find significant positive correlations between respect and all AQ dimensions: even if respect is only implicitly captured in the AQ guidelines (e.g. an argument is scored higher if it uses *appropriate language* which is by definition respectful), arguments of a higher AQ are also more respectful.

5.2. Results

In order to investigate whether the results would improve (especially for JUSTIFICATION and RESPECT, which positively correlate with AQ), as a next step we explicitly incorporate the AQ scores into the classifiers. We use the human-annotated dataset (*EuropolisAQ*) for these experiments, creating a new five-fold stratified split for each dimension. The amount of training data for this experiment is a lot smaller compared to the first experiment, so we use the same data augmentation method (see Table 13 in the appendix for training sizes of each dimension). We experiment with the following classifiers:

Boosted tree-ensemble: We either only train the classifier using the AQ scores (**AQ tree**) or on a combination of the linguistic features and the AQ scores (**feats+AQ tree**). We compare the results to the classifier that is based on the linguistic features alone (**feats tree**).

Roberta-base: We adapt the classification head of `roberta-base` and incorporate the AQ scores as features. We concatenate the `[CLS]` representation that is usually fed trough the classification head with the 4-dimensional vector containing the AQ scores and then apply the standard feed-forward layer adapting size of the initial transformation. We report the result on the original training set (**roberta-AQ**) and the augmented training data (**roberta-augment-AQ**). We make the simplifying assumption that the AQ scores do not change for the augmented data, and use scores from the corresponding annotated original comment.

The results are shown in Table 7: using AQ scores instead of linguistic features for the tree-based models leads to slight improvements (e.g. best performance for feats+AQ tree for JUSTIFICATION and INTERACTIVITY) and bring the largest improvements for RESPECT (+6% when using only AQ instead of features). They lead to a drop in performance in this setup for COMMON GOOD. Incorporating AQ scores into roberta does not improve but hurt the performance, for all dimensions.

	just	c.good	int	resp
feats tree	0.30	0.44	0.26	0.36
AQ tree	0.31	0.30	0.25	0.42
feats+AQ tree	0.32	0.39	0.27	0.38
roberta-base	0.50	0.57	0.67	0.35
roberta-AQ	0.24	0.33	0.15	0.29
roberta-augment	**0.77**	**0.89**	**0.83**	**0.79**
roberta-augment-AQ	0.54	0.38	0.32	0.49

Table 7: F1 macro score for the models with and without AQ scores, trained and tested on each DQ on *EuropolisAQ*: (just[ification], c[ommon].good, int[eractivity], resp[ect])

The adapted models evidently failed to incorporate the additional AQ knowledge, and this could be due to multiple reasons, which correspond to future work directions. The first is the strategy to incorporate the AQ scores: alternative options could be to provide the AQ scores in a textual form or to use attention to better combine the textual and numerical feature representations. Another potential reason for this negative result is the reliability of the AQ scores from our small-scale pilot annotation study; in this connection, the next step is clearly to gather more AQ annotations.

The general positive impact of the data augmentation method is replicated here (strong improvements in all DQ dimensions) , in a dataset that has an even smaller size than the original one.

6. Conclusion

With this work, we lay a foundation for empirical research on deliberative theory by enabling automatic prediction of deliberative quality. We enrich an existing dataset with expert annotation on deliberative quality in many ways: with data augmentation (which alleviates the small size and unbalanced classes of the dataset) and with features which encode linguistic properties and argument quality (the NLP counterpart of deliberative quality, at least in our assumption). Data augmentation turns out to be very successful, while the integration of argument quality is so far unsuccessful probably also due to the extremely challenging nature of its annotation on political science data. An additional contribution of our paper is our analysis of the most salient linguistic and lexical features picked up by the classifiers, which is of particular relevance for any NLP/machine learning study but in particular for one like ours that explicitly targets an interdisciplinary audience.

[8] $*$ marks statistical significance of $p < 0.001$.

Acknowledgments

The research reported in this paper has been funded by Bundesministerium für Bildung und Forschung (BMBF) through the project E-DELIB (Powering up e-deliberation: towards AI-supported moderation). We thank the anonymous reviewers for their valuable feedback. We would like to thank Marlène Gerber for giving us access to the Europolis dataset.

7. Bibliographical References

Ansari, G., Garg, M., and Saxena, C. (2021). Data augmentation for mental health classification on social media. *ArXiv*, abs/2112.10064.

Crossley, S. A., Kyle, K., and McNamara, D. S. (2017). Sentiment analysis and social cognition engine (seance): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior research methods*, 49(3):803–821.

Dai, X. and Adel, H. (2020). An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Fishkin, J. (1995). *The Voice of the People: Public Opinion and Democracy*. Yale University Press.

Fournier-Tombs, E. and Di Marzo Serugendo, G. (2019). Delibanalysis: Understanding the quality of online political discourse with machine learning. *Journal of Information Science*, 46:016555151987182, 09.

Fournier-Tombs, E. and MacKenzie, M. K. (2021). Big data and democratic speech: Predicting deliberative quality using machine learning techniques. *Methodological Innovations*, 14(2):20597991211010416.

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

Gretz, S., Friedman, R., Cohen-Karlik, E., Toledo, A., Lahav, D., Aharonov, R., and Slonim, N. (2020). A large-scale dataset for argument quality ranking: Construction and analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7805–7813, Apr.

Habermas, J. (2005). Concluding comments on empirical approaches to deliberative politics. *Acta politica*, 40(3):384–392.

Kyle, K., Crossley, S., and Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (taales): version 2.0. *Behavior research methods*, 50(3):1030–1046.

Kyle, K., Crossley, S. A., and Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, 18(2):154–170.

Lauscher, A., Ng, L., Napoles, C., and Tetreault, J. (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*, pages 4563–4574, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Mou, G., Li, Y., and Lee, K. (2021). Reducing and exploiting data augmentation noise through meta reweighting contrastive learning for text classification. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 876–887. IEEE.

Ng, L., Lauscher, A., Tetreault, J., and Napoles, C. (2020). Creating a domain-diverse corpus for theory-based argument quality assessment. In *Proceedings of the 7th Workshop on Argument Mining*, pages 117–126, Online, December. Association for Computational Linguistics.

Steenbergen, M., Baechtiger, A., Spörndli, M., and Steiner, J. (2003). Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1:21–48.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Thompson, D. F. (2008). Deliberative democratic theory and empirical political science. *Annual Review of Political Science*, 11(1):497–520.

Toledo, A., Gretz, S., Cohen-Karlik, E., Friedman, R., Venezian, E., Lahav, D., Jacovi, M., Aharonov, R., and Slonim, N. (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, Hong Kong, China, November. Association for Computational Linguistics.

Vecchi, E. M., Falk, N., Jundi, I., and Lapesa, G. (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, Online, August. Association for Computational Linguistics.

Wachsmuth, H. and Werner, T. (2020). Intrinsic quality assessment of arguments. In *Proceedings of the 28th International Conference on Compu-*

tational Linguistics, pages 6739–6745, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Wachsmuth, H., Naderi, N., Habernal, I., Hou, Y., Hirst, G., Gurevych, I., and Stein, B. (2017a). Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255, Vancouver, Canada, July. Association for Computational Linguistics.

Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T. A., Hirst, G., and Stein, B. (2017b). Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain, April. Association for Computational Linguistics.

Wei, J. and Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China, November. Association for Computational Linguistics.

Yi, M., Hou, L., Shang, L., Jiang, X., Liu, Q., and Ma, Z.-M. (2021). Reweighting augmented samples by minimizing the maximal expected loss. *arXiv preprint arXiv:2103.08933*.

8. Appendix

8.1. Discourse Quality and Argument Quality Annotations

Discourse Quality Annotation based on the DQI The following tables provide the excerpts from the guidelines for each of the four deliberative quality dimensions, taken from the codebook used to create the Europolis corpus (Gerber et al., 2018).

- Table 8 shows the guidelines for level of justification. We merged sophisticated justification (broad) and sophisticated justification (in depth).
- Table 9 depicts the guidelines for reference to common good. We merged the classes reference to common good in the utilitarian sense (e.g. EU) and in terms of the difference principle (e.g. solidarity
- Table 10 refers to interactivity. No labels were merged
- Table 11 shows the guidelines for respect. We dropped the class respect (balanced) as there were too few data points for that.

Argument Quality Table 12 shows the general description for each of the core sub-dimensions of Argument Quality, taken from the guidelines from Ng et al. (2020). The detailed guidelines that have been provided to our annotators as is, can be found under the following link `https://github.com/grammarly/gaqcorpus`.

8.2. Features

This section provides the details regarding the features briefly introduced in section 4.1 and employed in the experiments. Tables 14, 15 and 16 list all features names grouped by type, along with a short description and information on the values.

Lexical diversity (Table 14) These metrics are different variants of the type/token ratio, designed to be less sensitive to text length.
These features has been extracted with `TAALED` [9]. For more details refer to Kyle et al. (2021).

Lexical sophistication (Table 15) The metrics of lexical sophistication are computed based on word / co-occurrence information taken from existing reference corpora and word lists, e.g. the Corpus of Contemporary American English (COCA) or the (Averil Coxhead's) High-Incidence Academic Word List (AWL).

- *Word Frequency*: given a text, its word frequency value is calculated as the average of the frequencies of the words occurring in it, based on frequency estimates from different reference corpora (see above).

[9] `https://www.linguisticanalysistools.org/taaled.html`

No Justification	The speaker does not present any argument or only says that X should or should not be done, but no reason is given.
Inferior Justification	Here a reason Y is given why X should or should not be done, but no linkage is made between X and Y—the inference is incomplete or the argument is merely supported with illustrations.
Qualified Justification	A linkage is made why one should expect that X contributes to or detracts from Y. A single complete inference already qualifies for code 2.
Sophisticated Justification (broad)	At least two complete justifications are given, either two complete justifications for the same demand or complete justifications for two different demands.
Sophisticated Justification (in depth)	At least two complete justifications are given, either two complete justifications for the same demand or complete justifications for two different demands and discussed in depth.

Table 8: Annotation Guidelines for justification.

No reference	The speaker does not refer to benefits and costs at all.
Own country	Explicit statement concerning constituency or group interests (own country)
Explicit statement	Explicit statement in terms of a conception of the common good in utilitarian or collective terms (EU, Europe)
Explicit statement	Explicit statement in terms of the difference principle (solidarity, quality of life, global justice, etc.)

Table 9: Annotation Guidelines for COMMON GOOD.

- *Range indices*: given a text, its range indices are calculated as the average of document frequencies of the words occurring in it, estimated on reference corpora.
- *Mutual information*: uses the mutual information scores of academic bigrams, computed based on reference corpora.
- *Academic list indices* relative amount of academic words and n-grams using word lists as reference.
- *(Psycholinguistic) Word Information*: average of different psycholinguistic scores (e.g. concreteness, familiarity, imageability).
- *Semantic network*s: measures indicate how word forms are semantically related. More sophisticated texts contain words with fewer senses and words with more hypernyms (more subordinate terms).
- *Contextual distinctiveness* measures the diversity of contexts in which a word is encountered, e.g. "love" occurs in many different contexts, while the number of contexts where the word "bride" occurs is more restricted.

This set of features has been extracted with TAALES [10], see Kyle et al. (2018) for details.

Sentiment features (Table 16) The sentiment features rely on a number of pre-existing sentiment, social-positioning and cognition dictionaries (e.g. EmoLex) which serve as a look-up table.The features correspond to macro-feature component scores produced by PCA.
To extract the sentiment features, we use SEANCE [11]. The metrics and the retrieval of the feature components are described in Crossley et al. (2017).

8.3. Classification experiments

In what follows, we provide the implementation details for the classification models employed in our experiments and additional information about the size of the training sets of the second experiment.

- Gradient Boosted Trees: we use the XGBClassifier and the python package `xgboost` (`https://xgboost.readthedocs.io/en/stable/`. We use the default parameters and 100 estimators.
- Roberta: we tuned the following learning rates: 1e-5, 2e-5, 5e-5, 6e-5 on the validation sets. The hyperparameters for the final model are: learning rate: 5e-5, sequence length: 512 (captures most of the comments in full length), training batch size: 16. We use 3 GPUs of type *NVIDIA RTX A6000*.

Table 13 depicts the training size of the subset of Europolis, that has been annotated with AQ and the training size of the augmented data.

8.4. Model introspection

SHAP The following beeswarm plots combine feature importance and feature effects and shows the dis-

[10] `https://www.linguisticanalysistools.org/taales.html`

[11] `https://www.linguisticanalysistools.org/seance.html`

Negative	Negative (disrespectful) reference to other participants' arguments.
No reference	No reference to other participants' arguments.
Neutral	Neutral reference to other participants' arguments.
Positive	Positive (explicitly respectful) reference to other participants' arguments.

Table 10: Annotation Guidelines for INTERACTIVITY.

No Respect	This code is reserved for speeches in which there are only or predominantly negative statements about the groups.
Implicit Respect	No explicitly negative statements can be identified, but neither are there explicit positive statements.
Respect (balanced)	Both, positive and negative respect is equally expressed.
Explicit Respect	This code is assigned if there is at least one explicitly positive statement about the groups and either are negative statements completely absent or positive statements are clearly dominating the negative statements.

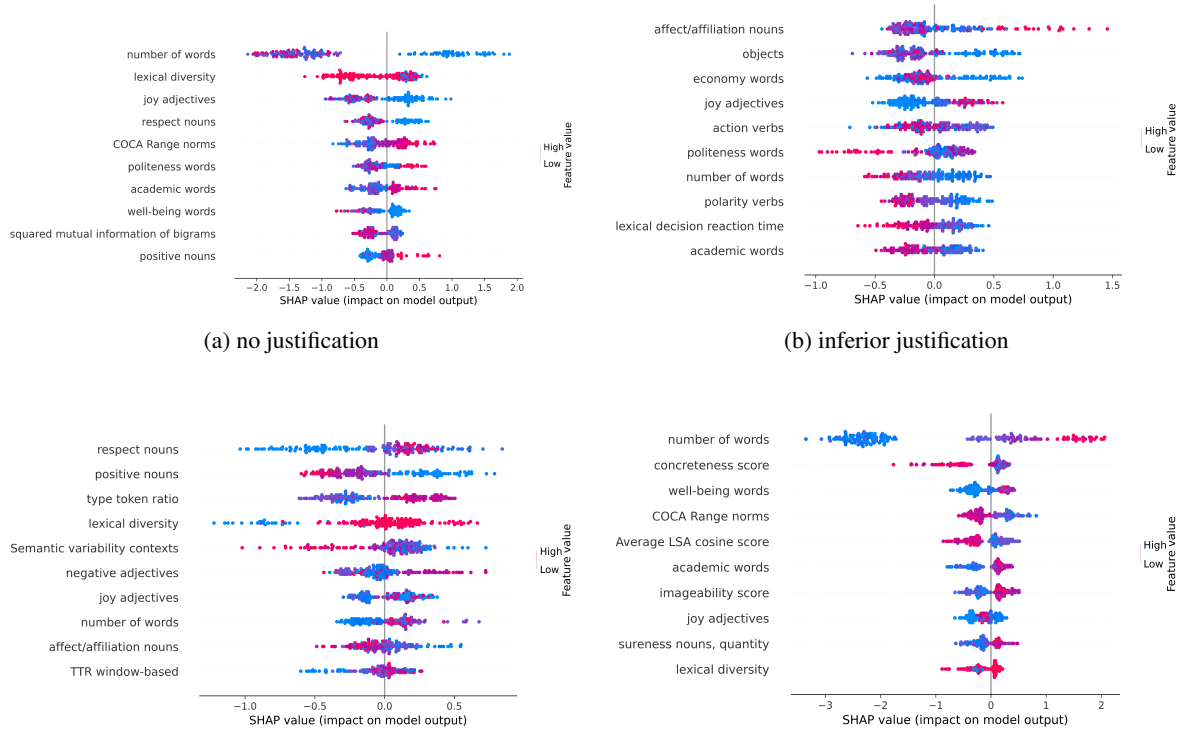
Table 11: Annotation Guidelines for RESPECT.

tribution of Shapley values for each feature over all instances for each DQ dimension. Figure 4 for justification, Figure 5 for common good, Figure 6 for interactivity and Figure 7 for respect. The features on the y-axis are ordered according to their importance. By looking at the position of the points on the x-axis we can see whether a feature has a positive or negative impact on the model output for a specific class. The color tells us whether the actual feature value is low (blue) or high (red). To compute the Shapley values and create the visualizations we used the python package `shap` (`https://github.com/slundberg/shap`)

LIG We used the python package `transformers-interpret` (`https://github.com/cdpierse/transformers-interpret#sequence-classification-explainer`) to compute numeric attributions for each instance in the test sets, for each DQ. Tables 17, 18, 19 and 20 show words that retrieved high positive attribution scores frequently (based on a list of all words with a positive attribution score of at least 0.1). The words are listed for each class separately and we compare content words (nouns, adjectives, verbs, adverbs) with other (function words and punctuation marks). Figures 2a and 2b and Figures 8b and 8b are visualizations of specific examples that are color coded: green indicates positive attributions values (positive impact on the probability of the predicted label) and red indicates negative attribution values. Words with higher absolute values are marked by color intensity.

Cogency	The argument includes acceptable justifications that are relevant to the point the author is making and that are sufficient to draw the author's conclusion.
Effectiveness	The way the argument is presented persuades you to agree with the author, e.g. the author changed your mind or affirmed a point you already agreed with.
Reasonableness	The argument contributes to the resolution of the given issue in a sufficient way that is acceptable to the target audience.
Overall	Judge the overall quality based on your ratings of cogency, effectiveness, and reasonableness. Also, take anything outside of these three traits that influences argument quality into account.

Table 12: General descriptions for each of the dimensions of Argument Quality taken from the original guidelines.



(a) no justification

(b) inferior justification

(c) qualified justification

(d) sophisticated justification

Figure 4: Beeswarm plot for JUSTIFICATION. The horizontal location shows whether the effect of that value is associated with a higher or lower prediction. Color shows whether that feature value is high (in red) or low (in blue).

	orig. train size	augm. train size
justification	279	968
common good	279	1,227
interactivity	279	514
respect	265	811

Table 13: Average training size obtained from *EuropolisAQ*: original training set vs. augmented

feature name	description	value
mtld_original_aw	computes type token ratio of increased word windows / segments	mean of all scores
mattr50_aw	Moving average type token ratio (50-word window)	mean of all scores
hdd42_aw	for each word type, compute the probability of encountering one of it's tokens in a random sample of 42 tokens, same range as type token ratio	mean of all scores

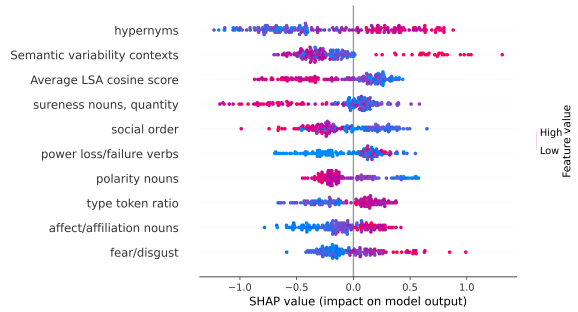
Table 14: Lexical diversity features: overview. Total number: 3.

feature name	feature type	description	value
COCA_spoken_Bigram_Frequency	N-gram	academic bigram frequency scores	mean of all scores
COCA_spoken_Frequency_AW	Word Frequency	frequency scores of words in spoken language	mean of all scores
COCA_spoken_Range_AW	Range indices	number of documents that the words occurs, domain: spoken language	mean of all scores
COCA_spoken_bi_MI2	mutual information	bigram association strength (mutual information squared), academic bigrams	mean of all scores
All_AWL_Normed	Academic list indices	number of academic words	relative amount of academic words
WN_Mean_Accuracy	Word Information	Average naming accuracy	mean of all scores
LD_Mean_Accuracy	Word Information	Average lexical decision accuracy	mean of all scores
LD_Mean_RT	Word Information	Average lexical decision accuracy	mean of all scores
MRC_Familiarity_AW	Word Information	unigram familiarity scores, MRC database	mean of all scores
MRC_Imageability_AW	Word Information	unigram imageability scores, MRC database	mean of all scores
Brysbaert_Concreteness_Combined_AW	Word Information	concreteness norms by Brysbaert et. al. (2013)	mean of all scores
McD_CD_AW	Contextual Distinctiveness	Co-occurrence probability of word with 500 highly frequent context lemmas (within 5 unigrams to the left and right of the target lemma)	Kullback-Leibler divergence relative entropy
Sem_D_AW	Contextual Distinctiveness	Semantic variability of contexts (1,000-word chunks of text) in which word occurs	Natural log of mean LSA cosine of similarity between contexts containing target words; reverses sign
content_poly	semantic networks	number of senses of content words	mean of all scores
hyper_verb_noun_Sav_Pav	semantic networks	hypernymy score for nouns and verbs, all senses and paths	mean of all scores

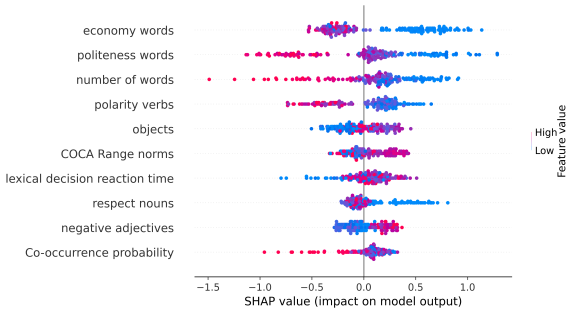
Table 15: Lexical sophistication features: overview. Total number: 16.

feature name	description
action_component	ought verbs, try verbs, travel verbs, descriptive action verbs
affect_friends_and_family_component	affect nouns, participant affect, kin noun, affiliation nouns
certainty_component	sureness nouns, quantity
economy_component	economy words
failure_component	power loss verbs, failure verbs
fear_and_digust_component	fear- / disgust- / negative nouns
joy_component	joy adjectives
negative_adjectives_component	negative adjectives
objects_component	objects
polarity_nouns_component	polarity nouns, aptitude nouns, pleasantness nouns
polarity_verbs_component	polarity verbs, aptitude verbs, pleasantness verbs
politeness_component	politeness nouns
positive_adjectives_component	positive adjectives
positive_nouns_component	positive nouns
positive_verbs_component	positive verbs
respect_component	respect nouns
social_order_component	ethic verbs, need verbs, rectitude words
trust_verbs_component	trust verbs, joy verbs, positive verbs
virtue_adverbs_component	hostility adverbs, rectitude gain adverbs, sureness adverbs
well_being_component	well-being words

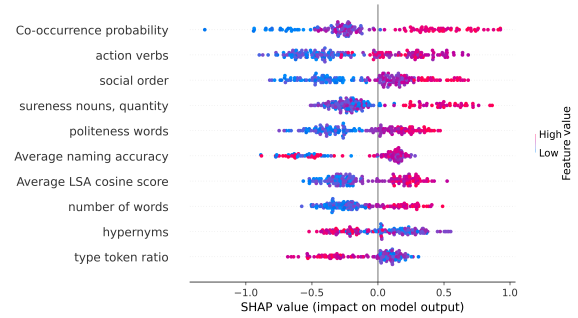
Table 16: Sentiment features: overview. Total number: 20



(a) no reference to common good

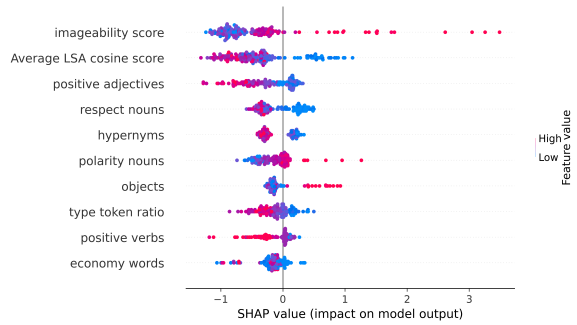


(b) referernce to own country

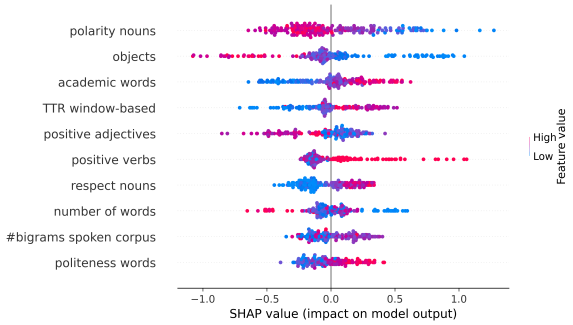


(c) reference to common good

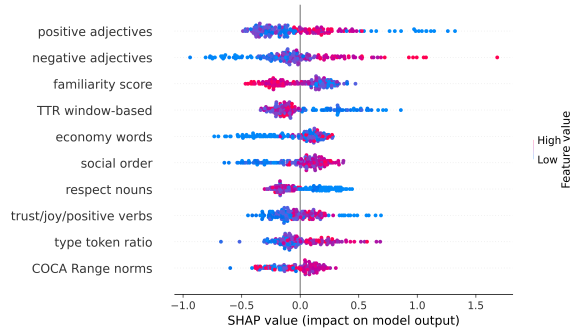
Figure 5: Beeswarm plot for COMMON GOOD. The horizontal location shows whether the effect of that value is associated with a higher or lower prediction. Color shows whether that feature value is high (in red) or low (in blue)



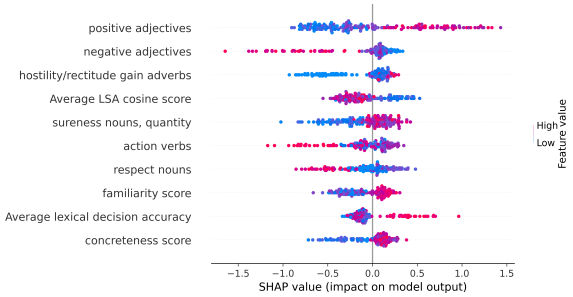
(a) negative reference



(b) no reference

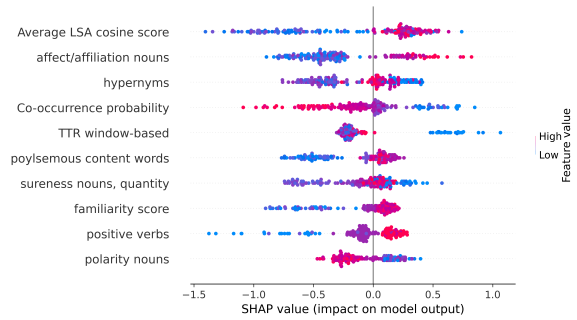


(c) neutral reference

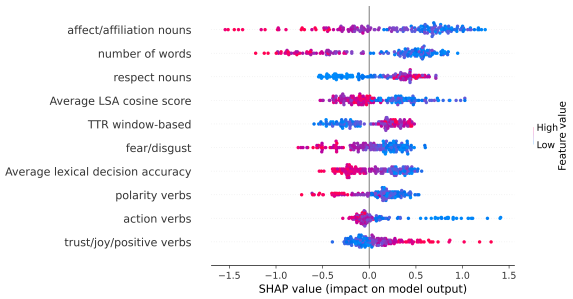


(d) positive reference

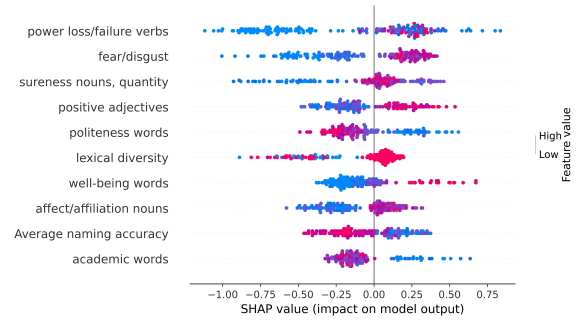
Figure 6: Beeswarm plot for INTERACTIVITY. The horizontal location shows whether the effect of that value is associated with a higher or lower prediction. Color shows whether that feature value is high (in red) or low (in blue)



(a) disresepectful



(b) implicit respect



(c) explicit respect

Figure 7: Beeswarm plot for RESPECT. The horizontal location shows whether the effect of that value is associated with a higher or lower prediction. Color shows whether that feature value is high (in red) or low (in blue).

class	content words	function words
inferior	not, country, can, have, is, countries, people, should, there, problem, be, so, EU, example, ”t”, culture, , think, problems	we, to, the, that, I, of, in, it, a, because, .
qualified	country, people, can, have, should, countries, not, think, Europe, must, be, problem, EU, is, problems, different, are, agree, money	because, to, if, we, the, that, of, I, for, a, .
no justification	is, are, question, immigrants, not, ”s”, ”t”, So, illegal, know, do, people, Germany, also, language	the, I, to, that, and, of, a, they, it, who, ., ?, :, -
sophisticated	country, countries, people, borders, immigration, money, should, must, cannot, population, think, companies, problem, pay, not, economic, misery, life, bad	because, I, a, of, to, our, and, the, in, who, .

Table 17: Most frequent words with high attributions for JUSTIFICATION. Model: roberta-augment

class	content words	function words
common good	Europe, European, EU, should, not, countries, is, can, global, think, agree, Well, Europeans, member, level, would, are, so	we, the, to, I, and, that, in, one, 27, us, .
own country	immigrants, immigration, country, people, is, countries, have, there, are, illegal, work, legal, know, language, ”s”, example, not, problem,	the, of, a, to, about, who, in, that, what, I, .
no reference	immigrants, country, immigration, problem, countries, Spain, France, Poland, are, Germany, work, know, Hungary, have, problems, illegal, more, people, not	in, to, the, I, from, each, their, our, of, all, .

Table 18: Most frequent words with high attributions for COMMON GOOD. Model: roberta-augment

class	content words	function words
neutral	immigrants, problem, immigration, there, have, illegal, example, are, is, not, ”s”, know, different, people, ”t”, do, question, come	to, the, that, I, a, of, in, about, for, this, ., ?
no reference	immigrants, immigration, illegal, have, people, example, country, not, is, countries, problem, are, say, come, should, be, borders	the, to, a, I, of, in, that, who, we, about, ., ?
negative	say, not, are, ”t”, can, have, , whole, world, money, put, children, ”s”, So, saying, nothing, doesn, is, EU	a, I, their, to, them, us, my, we, all, they, .
positive	agree, good, true, very, right, should, also, problem, immigration, think, is, example, important, said, always, ”s”, not, So, totally	I, with, Yes, that, the, this, because, like, it, you, .

Table 19: Most frequent words with high attributions for INTERACTIVITY. Model: roberta-augment

class	content words	function words
implicit	countries, country, immigration, people, problem, have, money, work, would, there, European, immigrants, solution, should, question, EU, policy, be, problems	to, the, a, this, in, I, we, of, that, for, .
disrespectful	language, are, have, integrate, bian, is, Poland, problem, ”s”, Germany, know, has, immigrants, here, then, region, whole	the, I, you, to, and, of, That, that, with, their, ., -
explicit	people, immigrants, would, life, illegal, live, problems, better, Europe, do, money, not, opinion, have, is, always, are, human, eat	to, these, the, in, that, of, them, if, I, we, .

Table 20: Most frequent words with high attributions for RESPECT. Model: roberta-augment

#s As far as the discussion is concerned ; we have been saying that within the EU , there are countries where the economic development is far lower than in other countries . That is why the EU funds could be invested in those less developed countries , in order to make their chances equal to everyone . That would mean that EU countries should contribute a part of the generated wealth and welfare to the poorer . #/s

(a) COMMON GOOD: example for *reference to common good*

#s Yes , two examples of values : I think in Western Europe , for example , I take the image of women or I take the point of view of revenge , which is clearly differently formed and translated into norms in the realm of Islam and also in their peoples than in ours . I do believe that there Europe and Islam are clearly different . #/s

(b) RESPECT: *disrespectful* contribution

Figure 8: Examples of Europolis with highlighted important words picked up by the best model: *green* words have a positive effect on the corresponding class and *red* words a negative effect.