LREC-COLING 2024

The First Workshop on Language-driven Deliberation Technology (DELITE2024)

Workshop Proceedings

Editors
Annette Hautli-Janisz, Gabriella Lapesa, Lucas Anastasiou,
Valentin Gold, Anna De Liddo, Chris Reed

20 May, 2024
Torino, Italia

Proceedings of the First Workshop on Language-driven Deliberation Technology (DELITE2024): Workshop Proceedings

ISBN 978-2-493814-14-2
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association
and the International Committee on Computational Linguistics

Preface

Deliberation is ubiquitous: from navigating divergent interests in everyday personal life to reaching consensus in the political decision making process, deliberation describes the communicative process by which a group of people exchange ideas, weigh different arguments, and ultimately reach mutual understanding. In recent years, deliberative processes have gained momentum and shown to improve everyday and political decision-making. For the first time, technological solutions are maturing to the point that they can be deployed to support deliberation. In this context, we want to establish the foundations for collecting and curating data for deliberation domains and for evaluating technology in deliberative settings.

The DELITE workshop provides a forum for presenting new advances in technology around deliberation by addressing researchers in Natural Language Processing, human-computer interaction, corpus linguistics, political science and philosophy, as well as stakeholders and domain experts involved in integrating such technology into decision-making processes.

Organizing Committee

Annette Hautli-Janisz (University of Passau)
Gabriella Lapesa (Lebniz Institute for the Social Sciences (GESIS), Köln, Heinrich-Heine University Düsseldorf)
Lucas Anastasiou (The Open University)
Valentin Gold (University of Göttingen)
Anna De Liddo (The Open University)
Chris Reed (University of Dundee)

Table of Contents

Workshop Program

Monday 20 May 2024

14:00–15:00 **Long papers**
Chair: Gabriella Lapesa

14:00–14:20 *AQuA – Combining Experts' and Non-Experts' Views To Assess Deliberation Quality in Online Discussions Using LLMs*
Maike Behrendt, Stefan Sylvius Wagner, Marc Ziegele, Lena Wilms, Anke Stoll, Dominique Heinbach and Stefan Harmeling

14:20–14:40 *A Unified LLM-KG Framework to Assist Fact-Checking in Public Deliberation*
Nikolaos Giarelis, Charalampos Mastrokostas and Nikos Karacapilidis

14:40–15:00 *Can Text Simplification Help to Increase the Acceptance of E-participation?*
Regina Stodden and Phillip Nguyen

15:00–15:20 **Short papers**
Chair: Lucas Anastasiou

15:00–15:05 *Pitfalls of Conversational LLMs on News Debiasing*
Ipek Baris Schlicht, Defne Altiok, Maryanne Taouk and Lucie Flek

15:05–15:10 *Integrating conflict prevention tools into deliberative democracy online platforms*
Sara Greco and Chiara Jermini

15:10–15:15 *A Hybrid Human-AI Approach for Argument Map Creation From Transcripts*
Lucas Anastasiou and Anna De Liddo

15:15–15:20 *Leveraging High-Precision Corpus Queries for Text Classification via Large Language Models*
Nathan Dykes, Stephanie Evert, Philipp Heinrich, Merlin Humml and Lutz Schröder

15:40–16:00 ***Poster session***

16:00–16:30 ***Coffee break***

16:30–17:00 ***Poster session (continued)***

Monday 20 May 2024 (continued)

17:00–18:00 *Panel*

Chair: Chris Reed (ARG-tech, University of Dundee)

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

The First Workshop on Language-driven Deliberation Technology (DELITE2024), pages 1–12

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.

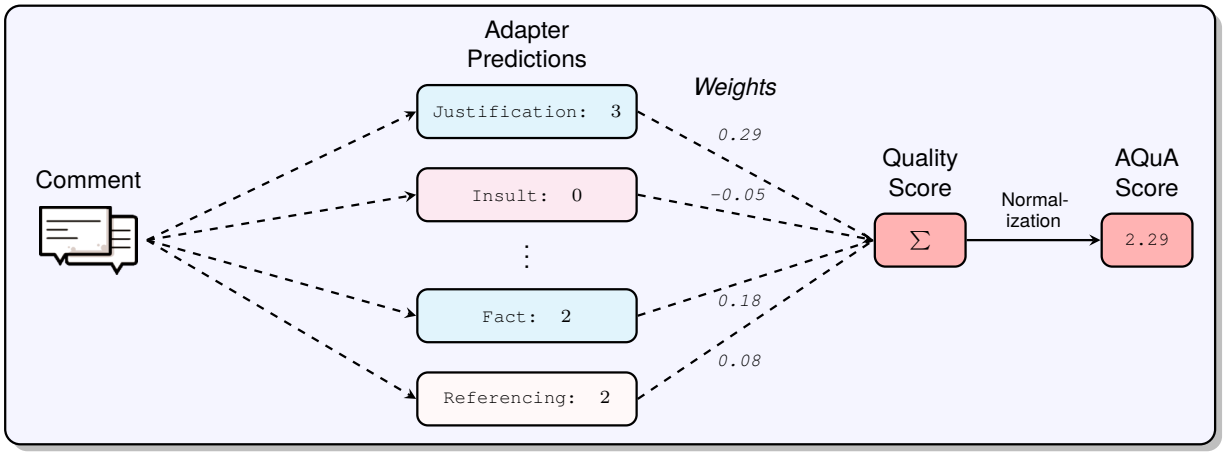


Figure 1: AQuA calculates a single score for deliberativeness from weighted adapter predictions on 20 different deliberative aspects. The adapter predictions are weighted by the correlation coefficients between each deliberative aspect and the perception of crowd workers about whether a comment is deliberative 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.

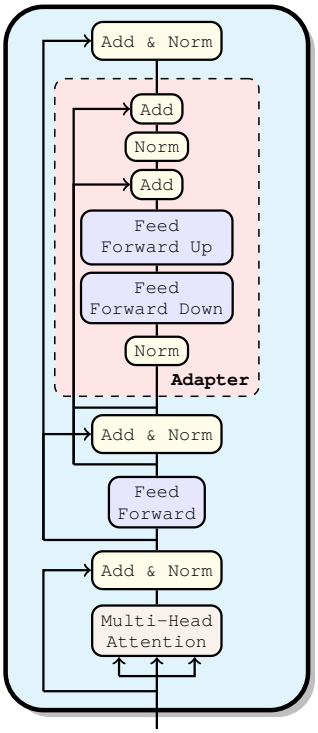


Figure 2: For the individual adapter predictions, we use a Transformer based model with adapter layers inserted after the feed forward layer of the Transformer as proposed by Pfeiffer et al. (2021).

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:

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

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),

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

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

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

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

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

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.

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

		German BERT	Multilingual BERT cased	Multilingual BERT 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

	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.

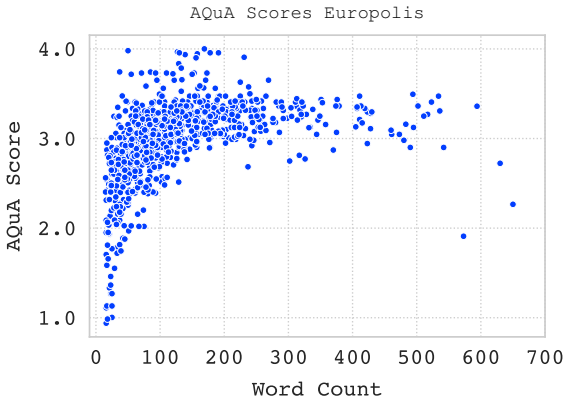


Figure 3: Europolis. AQuA scores (y-axis) vs the comment length (x-axis, word count) rule out that comment length alone is a factor for a high AQuA score.

[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.

A Unified LLM-KG Framework to Assist Fact-Checking in Public Deliberation

Nikolaos Giarelis, Charalampos Mastrokostas, Nikos Karacapilidis
Industrial Management and Information Systems Lab, MEAD
University of Patras, Rio Patras, Greece
{giarelis, cmastrokostas}@ac.upatras.gr, karacap@upatras.gr

Abstract

Fact-checking plays a crucial role in public deliberation by promoting transparency, accuracy, credibility, and accountability. Aiming to augment the efficiency and adoption of current public deliberation platforms, which mostly rely on the abilities of participants to meaningfully process and interpret the associated content, this paper explores the combination of deep learning and symbolic reasoning. Specifically, it proposes a framework that unifies the capabilities of Large Language Models (LLMs) and Knowledge Graphs (KGs), and reports on an experimental evaluation. This evaluation is conducted through a questionnaire asking users to assess a baseline LLM against the proposed framework, using a series of fact-checking metrics, namely readability, coverage, non-redundancy, and quality. The experimentation results are promising and confirm the potential of combining the capabilities of these two technologies in the context of public deliberation and digital democracy.

Keywords: Large Language Models, Knowledge Graphs, Fact Checking, Public Deliberation

1. Introduction

Public deliberation is a complex process that requires a close examination of diverse issues and listening to others' perspectives, aiming to shape and disclose the public judgement on what represents the common good (Bächtiger and Pedrini, 2010). To achieve its aims, it requires information about the topic under consideration, knowledge of the diverse associated elements and perspectives, as well as an understanding of the relationships among them. To ensure the accuracy and integrity of the above, stakeholders often need to engage with fact-checking processes. Admittedly, fact-checking plays a crucial role in public deliberation by promoting transparency, accuracy, credibility, and accountability; by providing citizens with verified information and countering misinformation, it contributes to a better-informed and constructive deliberation (Nyhan et al., 2020).

Digital solutions that support public deliberation and fact-checking rely mostly on the abilities of participants to meaningfully process and interpret the associated content (Kriplean et al., 2014). This may significantly limit the effectiveness of these solutions, especially in cases characterized by information overload and incomplete knowledge of participants on the subject under consideration. To thoroughly augment the effectiveness of digital deliberation platforms, and accordingly facilitate evidence-based collective decision making, we need to drastically enhance the synergy between human and machine reasoning that is supported by them (Kalampokis et al., 2023).

Current technological advancements from the areas of Large Language Models (LLMs) and Knowledge Graphs (KGs) may significantly contribute to the above issues. LLMs are advanced AI models that can complete complex reasoning tasks across various domains; their capabilities, along with the interaction with humans through intuitive chat interfaces, have led to their widespread adoption by the general public (Touvron et al., 2023). Due to their significant Natural Language Understanding (NLU) capabilities, LLMs can support and augment various reasoning tasks such as argument mining, summarization, and fact-checking, which are of paramount importance in an online large-scale public deliberation (Kriplean et al., 2014). LLMs are able to generalize and produce new information that is not part of their training knowledge. However, this knowledge is stored in a non-interpretable manner, due to their black-box architecture; moreover, their generalization capabilities can often lead to hallucinations, in cases where there is no proper context in their prompt.

On the other hand, KGs provide a flexible and powerful framework for organizing, integrating, and leveraging interconnected data and knowledge, enabling the extraction of valuable insights and informed decision making (Paulheim, 2017). KGs have structural knowledge that is stored in the form of accurate and interpretable domain-specific facts; however, they are unable to handle cases of missing or incomplete facts, and they do not possess any NLU capabilities.

In this paper, we explore the combination of the capabilities of these two technologies for fact-checking in public deliberation. The proposed approach addresses their limitations, in that it benefits from the contextual understanding of LLMs and the structured knowledge representation of KGs. Overall, the contribution of this work is twofold: (i) it proposes a unified LLM-KG framework that builds on the strengths of deep learning and symbolic reasoning; (ii) it reports on the results of its experimental evaluation, which was based on a questionnaire asking users to assess a baseline LLM against the proposed framework, using a series of fact-checking metrics.

The remainder of the paper is organized as follows. LLM- and KG-based approaches for fact checking are described in Section 2, pointing out benefits and limitations. The proposed framework for fact-

The First Workshop on Language-driven Deliberation Technology (DELITE2024), pages 13–19

checking, along with the experimental setup and results are presented in Section 3. Finally, concluding remarks and future research directions are outlined in Section 4.

2. LLM- and KG-based Approaches for Fact-checking

This section reports on works that utilize LLMs or KGs for fact-checking purposes, as well as on hybrid approaches that combine external knowledge bases with LLMs. To start with, ClaimsKG (Tchechmedjiev et al., 2019) is a knowledge graph containing verified claims; it is produced via a semi-automated pipeline, which gathers data from well-known fact-checking platforms, normalizes ratings and mentions of entities, resolves duplicate claims, and transfers the data into a specially designed model. This KG aims to support research in fact-checking, stance detection, and various topics related to analyzing societal debates. It enables structured queries of related metadata, such as truth values, authors or time release.

KLG-GAT (Zhu et al., 2021) is a fact checking model that relies on a "claim-evidence" knowledge graph architecture, which integrates unstructured evidence, structured knowledge triplets, and the claim itself. First, the most relevant sentences to the claim are selected from Wikipedia articles. Then, knowledge triplets relevant to both the claim and the retrieved evidence are extracted from a Wikipedia-based KG. After this selection, the triplets along with the claim and the retrieved evidence are fed into a BERT encoder to extract their semantic features, which are utilized as initial representations of the nodes in the graph. Additionally, a graph attention layer is utilized to perform reasoning over the graph and update the nodes' representations. Finally, the information between the claim, evidence, and knowledge (triplets) nodes in the graph are passed to a claim classification layer. The developers of this model report a 4% improvement in label accuracy on a fact-checking dataset compared to other state-of-the-art fact-checking models.

Wang et al. (2023) investigate the reasoning capabilities of LLM models by experimenting with debate-like conversations between OpenAI's ChatGPT (Brown et al., 2023) and users. Their goal is to determine whether the LLM can consistently maintain and defend its belief in the truth throughout a debate, without being misled by the user. The authors propose an evaluation framework that utilizes various benchmarks to evaluate the failure rate of ChatGPT across different types of reasoning tasks, including mathematics, logic and commonsense. Their results indicate that ChatGPT is susceptible to being misled into accepting falsehoods, revealing vulnerabilities not captured by traditional benchmarks. Finally, their work highlights potential risks associated with aligning models based on human feedback.

Castagna et al. (2024) present a comprehensive survey about argumentation-based chatbots and their abilities. Although their study focuses on earlier chatbot architectures, they also examine the benefits of using LLMs for argumentation purposes. The authors point out that despite their NLU capabilities, LLMs exhibit a set of limitations; specifically: (i) they struggle to explain their outputs even in the case of similar inputs, thus explainability is a vital requirement; (ii) they present factually incorrect information (hallucinations) based on false training data or mistakes in their reasoning process; (iii) they are characterized by weak reasoning skills, being unable to handle complex tasks; (iv) they generate toxic and/or offensive language in their outputs, due to such data being used during their training. Several techniques have been proposed in the literature according to the authors to mitigate these limitations, however these techniques do not fully address them.

Pan et al. (2024) propose various architectural combinations of LLMs with KGs from a theoretical perspective, as a means to complement the strengths and remedy the weaknesses of both. On one hand, they highlight that LLMs possess NLU and generalization capabilities, especially for unseen knowledge; however, they stress their inability of extracting facts due to their black-box design, and the potential for hallucinations (i.e., creating plausible yet incorrect facts). On the other hand, KGs contain easily interpretable and factually correct knowledge; nonetheless, this knowledge is often domain-specific and not timely updated, thus KG-based frameworks fail to generalize on new or unseen knowledge. Overall, the authors propose several frameworks which combine LLMs with KGs using various methodologies to improve the quality of fact-checking, question answering (QA) and reasoning. These methodologies include leveraging the retrieved KG facts as context for the LLM, injecting knowledge graph representations (i.e., graph embeddings) in the hidden LLM layers during training or during inference as special input tokens to the LLM. Research works that utilize these methodologies are proposed in (Wu et al., 2023) and (Yang et al., 2024).

Overall, the above works have a series of limitations. These include: (i) in the case of pure LLM-based approaches, the LLMs can easily lead to various issues (e.g., hallucinations); (ii) in the case of KG and LLM hybrid approaches, they are evaluated for simple and generic fact-checking and QA datasets, and have not been tested in a public deliberation setting, either by automatic or human evaluation; (iii) in the case of pure KG-based approaches, they only draw from generic fact-checking repositories (e.g., Wikipedia articles, fact-checking websites, etc.), which often do not contain domain-specific facts that are important for verifying user claims. The KG-based approaches also use them as static factual repositories, with no provision to dynamically update them in the future.

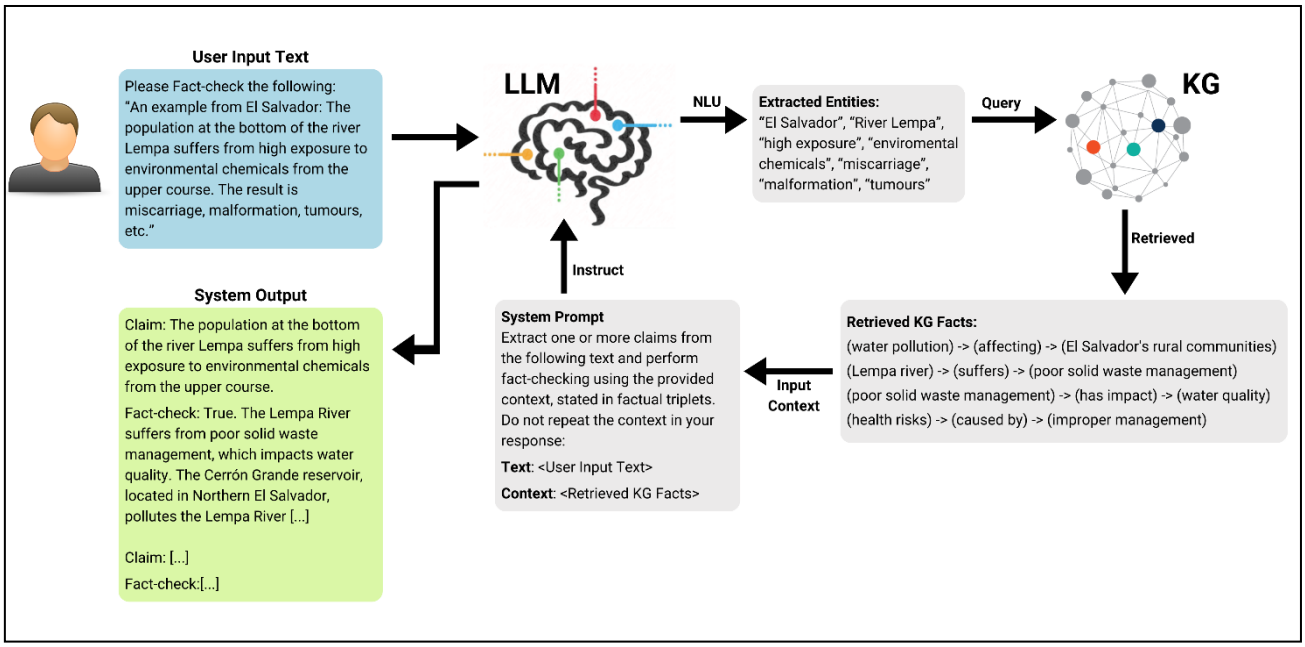


Figure 1. The proposed unified LLM - KG framework.

3. The Proposed Fact-checking Assistant

Aiming to build on the strengths of LLMs and KGs, while at the same time addressing their limitations and taking into account issues reported in the previous section, we propose a unified LLM-KG framework for fact-checking in public deliberation. The proposed solution combines state-of-the-art LLM technology (i.e., ChatGPT-3.5) with a dynamically updated KG. Specifically, it utilizes a prompt that instructs the LLM to perform fact-checking, while retrieving appropriate facts from the KG (Figure 1). These facts act as contextual sources for the LLM, improving the quality and validity of its responses. Furthermore, when confronted with lack of contextual knowledge, the proposed framework utilizes the LLM to introduce facts from verified sources back into the KG.

3.1 Our Approach

The first aspect of our approach concerns the construction of the KG that is used to verify the extracted claims and present supporting evidence in a public deliberation setting. For our KG, we store entities found in claims (e.g., sentence subjects or objects), using a generic entity node type. These are connected by edges that have as relationship type the predicate that connects them in text. For instance, the sentence "*EU nationals travel freely between European countries*" is transformed to the triplet: `(EU_Nationals)` → `(travel_freely)` → `(European_countries)`. In addition, some uniqueness constraints are placed to ensure that we avoid the storage of duplicated nodes and edges. Newly introduced facts with new edges (predicates) will further solidify the connections between existing node entities.

Initially, when the constraints are defined, the graph is empty; new entities are added to the KG using the following procedure, which is also used to update the KG in case of missing facts:

1. For each user input text, the proposed solution adds LLM-extracted evidence to the KG, using one of the following two ways:
   a) If the input text contains claims that point to existing studies (e.g., through a URL or a typical citation form, such as DOI), these are extracted by the LLM component. Then, the proposed framework crawls the URLs and extracts plaintext and text from the associated .pdfs (e.g., technical reports, scientific studies).
   b) If there are no such studies, we use the Google Search API to extract the first page of the top web search results that are related to the user claims. We filter these results to prioritize organizational (.org), governmental (.gov), or European union sites (.eu). These sites contain validated technical reports, usually in .pdf form.
2. In any of the above ways, these textual sources are parsed using the LLM's NLU capabilities to extract entities that are relevant to the user input text.
3. These LLM-extracted entities are then used to generate relevant triplets to be stored in the KG.

By using this process, we can build a dynamic KG that is updated with supporting evidence extracted from multiple data sources. This also helps to address the cold start problem, where initially there are no (or a limited number of) facts in the KG.

When a user intends to perform fact-checking, the proposed solution facilitates the following process (Figure 1):

- Firstly, the user requests fact-checking on a specific text.
- Secondly, the unified framework triggers its LLM component to extract important entities from the input using its NLU capabilities.
- Thirdly, these entities are used to query the KG; the KG returns a series of triplets, stored as `(subject)` → `(predicate)` → `(object)`, which contain related evidence.
- Finally, these triplets are used as contextual input to the LLM that factually enhances its prompt, aiming to generate a more appropriate answer.

3.2 Experimental Evaluation

For our experimentations, we utilized two publicly available deliberations (one about the export of hazardous chemicals by the EU, and another about the cultivation or import of Genetically Modified Organisms), both retrieved from the *"Have Your Say"* platform of European Commission (https://ec.europa.eu/info/law/better-regulation/have-your-say_en). From these deliberations, we created five example scenarios, where the LLM was prompted to perform fact-checking through two different setups. In both setups, the user's input text is placed into a system prompt, where the LLM is first asked to extract one or more claims and then check their validity. In the first setup, the baseline LLM is instructed to perform fact-checking using its internal knowledge, while in the second one the prompt is augmented with factual triplets retrieved from the KG. The LLM is asked to use only these triplets to perform fact-checking. In addition, it is instructed not to repeat the triplets verbatim in the generated answer.

A representative example of our experiments, including the user input, the exact system prompts, and the corresponding outputs, appears in the Appendix (all five examples can be found at https://forms.gle/GNZaZGXWk4PLsQch7). As shown in this example, the LLM fails to accurately capture the necessary context for fact-checking when it relies solely on its internal knowledge. As highlighted by the red text, the model constantly refers to the lack of concrete data. On the contrary, the proposed LLM-KG unified approach manages to overcome this problem and produce its response based on the provided factual knowledge, as highlighted by the green text.

In the literature, LLMs are usually evaluated using traditional metrics such as precision, recall and F1. However, this type of evaluation only counts the number of correct examples predicted by a model and does not assess qualitative aspects of the LLM-generated text from a human perspective. To evaluate the quality of the proposed examples, our experiments focus on human evaluation metrics that have been recently proposed for LLM-based fact-checking (Zhang and Gao, 2023). These metrics are:

1. **Readability**: The generated text is well-written and the provided explanations are clear.
2. **Coverage**: All important points for fact-checking are explained and appropriate reasoning is provided.
3. **Non-Redundancy**: The generated text provides relevant information as to understand the claims and fact-check it, while not repeating the same information.
4. **Quality**: The overall quality of the generated text.

These metrics are measured using the Likert scale (1-5). To measure the public opinion about the two different LLM setups, we developed a questionnaire including all five example scenarios (hosted at (https://forms.gle/GNZaZGXWk4PLsQch7). Participants were asked to rate each of the two setups using the above metrics, based on their subjective opinion. These metrics enable participants to assess the fact-checking capabilities of the two experimental setups. In the questionnaire, the ChatGPT-3.5 without context is labeled as "LLM-A", whereas the ChatGPT-3.5 with evidence-based KG context is labeled as "LLM-B". To avoid potential positive bias from respondents towards the proposed unified KG-LLM framework, we did not disclose the architectural difference between these two setups, thus allowing participants to comparatively assess them as they see fit.

	ChatGPT-3.5 (Baseline)	**ChatGPT-3.5 (KG-enhanced context)**
Readability	Max: 4.28 **Avg: 4.12** Min: 3.88	Max: 4.36 **Avg: 4.18** Min: 4.04
Coverage	Max: 4.04 **Avg: 3.55** Min: 3.16	Max: 4.20 **Avg: 3.62** Min: 3.32
Non-Redundancy	Max: 3.48 **Avg: 3.30** Min: 3.12	Max: 4.16 **Avg: 3.91** Min: 3.72
Quality	Max: 3.88 **Avg: 3.53** Min: 3.28	Max: 4.00 **Avg: 3.54** Min: 3.32

Table 1. Questionnaire results using the human evaluation metrics.

The *min*, *average*, and *max* values derived from the five examples for each metric are reported in Table 1 (for both setups). As shown, the proposed LLM-KG framework obtained better average scores with respect to *Readability* and *Coverage*, as well as a significantly improved average score with respect to *Non-Redundancy*. As far as the *Quality* metric is concerned, its average score was similar to the Baseline LLM.

Regarding the demographics of our study, we received responses from 25 participants. Most of

them belong in the “25-34 years old” (64%) age group. Our sample portrayed an imbalanced gender identity distribution (36% female and 64% male). In terms of education levels, most of the participants hold a master’s degree (40%), while all other educational levels had a relatively balanced representation. At the same time, most of our respondents are fluent in English (68%). Finally, it is noted that a considerable number of respondents declared themselves to be “very comfortable” (24%) or “somewhat comfortable” (32%) using LLMs (e.g., ChatGPT), with 28% of participants stating being “neutral” towards LLMs.

4. Discussion

This work proposes a unified LLM-KG framework to assist fact-checking in public deliberation platforms. It contributes to the mitigation of the hallucination and indecisiveness issues currently exhibited by LLMs, while improving transparency through factual context provided by KGs. The reported experimentation results are promising and confirm the potential of combining the capabilities of these two technologies in the context of public deliberation and digital democracy (Giarelis et al., 2023). In any case, we acknowledge a set of limitations and issues that require further attention:

- In our research, we used ChatGPT-3.5 as a state-of-the-art, yet proprietary, LLM; open-source LLMs, which have already reached or surpassed its performance (Jiang et al., 2024), have to be also considered.
- More experiments are needed to further validate the outcomes of our study. These may involve additional examples from various deliberation contexts, as well as a larger sample of participants from different backgrounds. They could also include quantitative experiments, which will measure the number of correct answers on various fact-checking datasets, using established metrics such as the F1 score.

Future research directions include:

- The use of open-source LLMs, for various reasons, including: (i) user privacy and EU regulatory concerns; user data are collected in OpenAI servers hosted in the US, with no GDPR compliance; (ii) reducing costs; these LLMs can be self-hosted, thus implying no additional costs; (iii) model explainability; since ChatGPT is a proprietary commercial product, its model weights cannot be accessed, thus no explainability methods can be applied.
- Since LLMs are directly linked to the quality of their generated text, future work needs to further incorporate human evaluation from diverse and large samples of participants.
- Work towards enhancing the proposed framework by revealing to the user the links pointing to the source material, aiming to improve users’ trust in automated fact-checking.
- The integration of diverse domain-specific knowledge from various data sources to the KG, thus facilitating the generalization of the proposed approach across multiple domains.

5. Bibliographical References

André Bächtiger and Seraina Pedrini. 2010. Dissecting deliberative democracy. A review of theoretical concepts and empirical findings. *Political discussion in modern democracies: A comparative perspective*:9–25.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, et al. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Federico Castagna, Nadin Kokciyan, Isabel Sassoon, Simon Parsons, and Elizabeth Sklar. 2024. Computational Argumentation-based Chatbots: a Survey. arXiv:2401.03454 [cs].

Nikolaos Giarelis, Nikos Karacapilidis, Georgios Kournetas, and Ilias Siachos. 2023. Investigating Trust in the incorporation of NLP applications in Digital Democracy Platforms. In *EGOV-CeDEM-ePart 2023*, Budapest, Hungary.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, et al. 2024. Mixtral of Experts. arXiv:2401.04088 [cs].

Evangelos Kalampokis, Nikos Karacapilidis, Dimitris Tsakalidis, and Konstantinos Tarabanis. 2023. Understanding the Use of Emerging Technologies in the Public Sector: A Review of Horizon 2020 Projects. *Digital Government: Research and Practice*, 4(1):4:1-4:28.

Travis Kriplean, Caitlin Bonnar, Alan Borning, Bo Kinney, and Brian Gill. 2014. Integrating on-demand fact-checking with public dialogue. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 1188–1199, New York, NY, USA. Association for Computing Machinery.

Brendan Nyhan, Ethan Porter, Jason Reifler, and Thomas J. Wood. 2020. Taking Fact-Checks Literally But Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability. Political Behavior, 42(3):939–960.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering*:1–20.

Heiko Paulheim. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, 8(3):489–508.

Andon Tchechmedjiev, Pavlos Fafalios, Katarina Boland, Malo Gasquet, Matthäus Zloch, Benjamin Zapilko, Stefan Dietze, and Konstantin Todorov. 2019. ClaimsKG: A Knowledge Graph of Fact-Checked Claims. In Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtěch Svátek, Isabel Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon, editors, *The Semantic Web – ISWC 2019*, pages 309–324, Cham. Springer International Publishing.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs].

Boshi Wang, Xiang Yue, and Huan Sun. 2023. Can ChatGPT Defend its Belief in Truth? Evaluating LLM Reasoning via Debate. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11865–11881, Singapore. Association for Computational Linguistics.

Yike Wu, Nan Hu, Sheng Bi, Guilin Qi, Jie Ren, Anhuan Xie, and Wei Song. 2023. Retrieve-Rewrite-Answer: A KG-to-Text Enhanced LLMs Framework for Knowledge Graph Question Answering. arXiv:2309.11206 [cs].

Linyao Yang, Hongyang Chen, Zhao Li, Xiao Ding, and Xindong Wu. 2024. Give Us the Facts: Enhancing Large Language Models with Knowledge Graphs for Fact-aware Language Modeling. *IEEE Transactions on Knowledge and Data Engineering*:1–20.

Xuan Zhang and Wei Gao. 2023. Towards LLM-based Fact Verification on News Claims with a Hierarchical Step-by-Step Prompting Method. arXiv:2310.00305 [cs].

Biru Zhu, Xingyao Zhang, Ming Gu, and Yangdong Deng. 2021. Knowledge Enhanced Fact Checking and Verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3132–3143.

Appendix: Fact-checking System Prompts and Responses

<User Input Text>: "An example from El Salvador: The population at the bottom of the river Lempa suffers from high exposure to environmental chemicals from the upper course. The result is miscarriage, malformation, tumours, etc."

<KG Facts>: "
(water pollution) → (affecting) → (El Salvador's rural communities)
(Lempa river) → (suffers) → (poor solid waste management)
(poor solid waste management) → (has impact) → (water quality)
(Cerrón Grande reservoir) → (pollutes) → (Lempa River)
(Cerrón Grande reservoir) → (is located) → (Northern El Salvador)
(Lempa river) → (receives pollution) → (Guatemala)
(Lempa river) → (receives pollution) → (Honduras)
(health risks) → (caused by) → (improper management)"

System Prompt (LLM without KG context)	**System Prompt (LLM + KG context)**
Extract one or more claims from the following text and perform fact-checking using your internal knowledge as a Large Language Model: Text: **<User Input Text>**	Extract one or more claims from the following text and perform fact-checking using the provided context, stated in factual triplets. Do not repeat the context in your response: Text: **<User Input Text>** Context: **<KG Facts>**
LLM Response (ChatGPT-3.5)	**LLM Response (ChatGPT-3.5)**
Claim: The population at the bottom of the river Lempa suffers from high exposure to environmental chemicals from the upper course, resulting in miscarriages, malformations, tumors, etc. Fact-check: **1. Exposure to Environmental Chemicals:** It's plausible that populations living downstream of industrial or agricultural areas could be exposed to environmental chemicals carried by the river from the upper course. This can happen through runoff or industrial discharge. Without specific data on the types and levels of chemicals present, it's challenging to verify the extent of exposure. **2. Health Effects - Miscarriages, Malformations, Tumors:** It's well-documented that exposure to certain environmental chemicals can increase the risk of adverse health effects, including miscarriages, birth defects (malformations), and tumors. However, linking specific health outcomes directly to the river Lempa without detailed epidemiological studies would be speculative. **3. Specificity of Location (El Salvador - River Lempa):** The River Lempa is indeed a significant waterway in El Salvador, flowing through various regions with different levels of industrial and agricultural activity. However, without specific data on pollution sources and their impact on downstream communities, it's challenging to confirm the exact situation described. Overall, while the claim aligns with known possibilities regarding environmental pollution and health risks, without specific data or studies confirming the situation in the River Lempa region, it remains unverified. Further investigation and research would be necessary to confirm the extent of the issue described.	Claim: The population at the bottom of the river Lempa suffers from high exposure to environmental chemicals from the upper course. Fact-check: True. The Lempa River suffers from poor solid waste management, which impacts water quality. The Cerrón Grande reservoir, located in Northern El Salvador, pollutes the Lempa River. Additionally, the river receives pollution from Guatemala and Honduras. This poor management contributes to health risks in El Salvador's rural communities, which includes high exposure to environmental chemicals. Claim: The result of high exposure to environmental chemicals in the river Lempa is miscarriage, malformation, tumors, etc. Fact-check: Partially true. While it's affirmed that the population at the bottom of the Lempa River is exposed to high environmental chemicals due to pollution, specific health impacts such as miscarriage, malformation, and tumors are not directly mentioned in the provided context. However, it's reasonable to infer that these health issues could occur as consequences of exposure to environmental chemicals, given the known risks associated with such exposure.

Can Text Simplification Help to Increase the Acceptance of E-Participation?

Regina Stodden[1] and Phillip Nguyen[2]
[1] Department of Computational Linguistics,
[2] Department of Business Administration
Heinrich Heine University Düsseldorf, Germany
regina.stodden@hhu.de, phillip.nguyen@hhu.de

Abstract

This study investigated the effect of text simplification (with and without artificial intelligence support) and the role of participants (author or reader) on the acceptance of e-participation processes. Therefore, a near-realistic experimental study with 276 participants was conducted simulating a participatory budgeting process. The results of our study show, on the one hand, that text simplification and the role of participants has no direct influence on the intention to use e-participation. Although a higher level of participation cannot be achieved by text simplification, our results also show that no negative consequences for usage intention can be expected from text simplification. On the other hand, the results show that people with reading and writing difficulties prefer text simplification for proposals in e-participation.

Keywords: text simplification, technology acceptance, participatory budgeting

1. Introduction

In light of current developments, such as the increasing complexity of public decision-making processes, the growing pluralization of society, or the increasing loss of trust in the competence of public decision-makers, digital forms of citizen participation are becoming an important component of government action (Panopoulou et al., 2014). Citizens are able to influence decisions or processes by submitting ideas, voting, and in a variety of other ways. One of the key success factors and goals of these online deliberative platforms is to engage a diverse group of citizens and achieve high levels of participation.

However, for certain groups, such as migrants or people with reading disabilities, the complexity of the procedures and texts can be a major barrier to participation (Zepic et al., 2017). In contrast, the simpler the platform texts, the easier they are to understand for people with and without reading difficulties (Gutermuth, 2019).

Moreover, legal requirements for inclusion ensure that accessibility is a central requirement and challenge for processes and procedures in the public sector (Ferri and Favalli, 2018). According to the recommendations of the European Standard for Digital Accessibility (European Telecommunications Standards Institute, 2021), public authorities should provide readable and understandable content (e.g., in plain language) on their websites to make them accessible to people with reading and writing deficits, e.g., people with cognitive limitations, low literacy skills or low language skills in a foreign language.

However, on many government websites, only a summary of the standard website is provided in plain language, and other related websites, such as those of citizen participation projects, are often written only in standard language (Asghari et al., 2023), which is difficult to read for people with reading and writing deficits. As a result, having plain language on online deliberative platforms is critical for both goal attainment and legal purposes.

Here, some natural language processing tasks (supported by artificial intelligence) could improve the accessibility for people with reading and writing difficulties in civic participation projects: For example, automatic grammar error or spelling correction (see Bryant et al. 2023) could make participants with low literacy skills more confident while publishing their proposals, text summarization (see El-Kassas et al. 2021) could help to reduce the content of the projects, so that people with reading deficits are not overwhelmed with the amount of data, or text simplification (see Alva-Manchego et al. 2020) could enhance the readability and comprehensibility of texts of citizen participation projects, with the result that people with reading deficits could understand the content and overall enable more people to participate in a project. This may facilitate their participation and improve the use of these platforms and thus their acceptance and success.

However, there is a lack of empirical studies dealing with the possible effects of text simplification in online deliberation processes such as e-participation.

Therefore, in our near-realistic experimental study, we first exploratively investigate attitudes towards a natural language processing (NLP) task,

The First Workshop on Language-driven Deliberation Technology (DELITE2024), pages 20–32

i.e. text simplification, in the context of a citizen participation project, i.e. participatory budgeting. Second, we analyze the concrete causal effect of text simplification on the acceptance of a citizen participation project and thus on the intention to participate in a citizen participation project.

In doing so, the study contributes to the discourse on the use of NLP in deliberative processes and increases knowledge about the possible effects of its use. On a practical level, the study helps to gain insight into the use of NLP and evaluate whether it can facilitate the control and revision of proposals in digital collaboration processes and reduce the costs of the processes.

In the remainder of this paper, we first outline some theoretical background and develop our hypotheses. We then present our research design before presenting the results of the readability study and the experimental study.

2. Theoretical Background & Related Work

2.1. E-participation

E-participation is the digitized form of citizen participation. In the fields of open government and e-government, e-participation has been widely discussed as an option for the growing demand for innovative methods to involve citizens and their opinions in the decision-making processes of public authorities (Simonofski et al., 2017).

Examples of e-participation include participatory budgeting, where citizens can vote on a set budget, or consultative procedures, where citizens can submit proposals for specific construction projects. While the use of information and communication technologies offers a number of opportunities to make processes more inclusive, e-participation is also usually associated with a number of technology-related challenges for certain groups of the population. Among other things, e-participation has the potential to open up new target groups by overcoming spatial and temporal limitations, whereas the participation of senior citizens or technology-skeptical citizens may decrease. However, since e-participation is also intended to make a significant contribution to optimizing democratic processes, accessibility is a key factor in ensuring that e-participation is accepted and used by the public.

2.2. Factors for Acceptance of e-participation

A multitude of potential influencing factors for acceptance of e-participation can be identified. For example, Naranjo Zolotov et al. (2018) found *perceived usefulness* to be one of the most predictive factors for the intention to use e-participation. *perceived ease of use* in turn is a major factor for perceived usefulness of e-participation.

The ease of use of e-participation platforms in terms of accessibility depends on the usability, perceptibility, operability, as well as the comprehensibility and readability (Vollenwyder et al., 2018).

The information overload on e-participation platforms can hinder the usability. Romberg and Escher (2023) summarized research on how to approach this problem using NLP methods. Following them, AI or NLP tools have been proposed to identify double proposals (see e.g., Yang et al. 2006), group proposals with topic modeling strategies (see e.g., Hagen et al. 2015), summarize the proposals for a shorter and faster readable overview (see e.g., Arana-Catania et al. 2021), or produce in-depth analysis (e.g., identification of argumentation structures in the proposals; see e.g., Liebeck et al. 2016).

2.3. Text Simplification & Plain Language & People with Reading and Writing Difficulties

In Terms of accessibility, the importance of the comprehensibility of the procedures, contents, and outcomes becomes apparent in the course of the discussion on diversity and equal opportunities for all citizens. For this purpose, it is necessary that every citizen is able to understand the procedural steps and contributions. For people with reading and writing deficits in the language of the e-participation project, the use of plain language is a key requirement to be able to comprehend the project. In this respect, on the one hand, the use of digital processes can actually help accessibility with respect to overcome language barriers. On the other hand, the digital approach (compared to a face-to-face approach) can also be a hindrance, e.g. people with insufficient computer skills may not be able to use sophisticated e-participation platforms (Zepic et al., 2017).

In general, Gutermuth (2019) has already shown that simplifying the instructions of citizen participation processes can help different groups of people to understand the instructions better. They measured the reading speed, the recall of the content, the eye movements during reading and much more of an instruction of a citizen participation project in Easy German (“Leichte Sprache”), Plain German (“Einfache Sprache”) and standard language by elderly people, people with immigrant background, people with cognitive impairment and a control group (students). The study discovered that all groups i) read the text faster when it was simpler, ii) had a better understanding of the easier versions

and iii) recalled the simpler text more effectively. The simpler the version, the more strong the effect. Therefore simple texts seem to be helpful for a wide range of people.

Although texts such as the instruction of the procedure can already be simplified and reviewed in advance (as shown by Gutermuth 2019), a simplification of the submitted citizen contributions' (hereafter proposals) requires a disproportionately higher effort. Considering the amount of content produced in citizen participation projects, a trained translator might be overwhelmed during the manual simplification process of the majority of proposals. However, automatic simplification of texts is a potential solution to reduce this effort. Text simplification (TS) is an NLP task that aims to automatically make complex texts more accessible by editing their wording and syntax, while preserving the original meaning of the text (Alva-Manchego et al., 2020).

In contrast to the previous study, Johnson et al. (2015) analyze the effect of the language of participants (and not the instructors) in online communities. Johnson et al. (2015) show that participants in an online community have a greater influence on other participants when their vocabulary is simpler, more readable and has a positive sentiment. In order to give all participants the opportunity to write more clearly and to read more readable texts, the simplification of these texts seems to be a relevant option.

However, depending on individual preferences, people without reading and writing deficits may perceive easy-to-read texts as less favorable and may be less satisfied with these texts than with standard texts (Karreman et al., 2007; Schmutz et al., 2019). Vollenwyder et al. (2018) extend the research of Karreman et al. (2007); Schmutz et al. (2019) by investigating whether these unwanted side-effects of people without reading and writing deficits can be overcome when the original (standard) text and the simplified version are presented in parallel. And indeed their results show that the positive effects of the simplified texts for the people with reading and writing deficits still remain, while the negative impact of people without reading and writing deficits disappears as they can still read the original content.

3. Hypotheses Building

Following the research previously presented, we are now building our hypothesis for our study.

The work of Gutermuth (2019) has already shown that instructions in standard German are difficult to read in online deliberation processes. In contrast to the instructions, the proposals are user-generated texts. This means that they are written by citizens and are not proofread before publication, hence, they might contain a high amount of ungrammatical sentences, and out-of-vocabulary words (Baldwin et al., 2013), which are difficult to process for people with reading deficits. Friess et al. (2017) also argue that the proposals of deliberation processes are also more difficult to understand than other (user-generated) texts because they contain many specific terminologies, emotions, arguments, and references to other proposals.

Further, the participants who write proposals in online deliberation processes are often academics (Schäfer and Schoen, 2013) who tend to write long and complex sentences. On the other hand, highly informal language with numerous errors and unusual features, which is common in digital participation processes (Parycek et al., 2014), can make the text also difficult to understand. Following this, we build H1:

> **H1**: Proposals of online e-participation processes are difficult to read.

Following the previous named characteristics of proposals of deliberation processes, people need the following skills to understand the proposals, e.g., comprehension of complex argumentation, comprehension of specific terminology (including regional and technical terms), comprehension of cohesive texts (e.g., reciprocal content and argumentation), and comprehension of sentences with complex structures (e.g., long sentences with many clauses) (Stodden, 2021). Following the self-assessment grid of the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2020) the acquisition of these skills corresponds to a CEFR level of B2. Bock (2015) also argues that people with reading and writing deficits have only limited access to political participation because they have fewer communication capabilities, e.g., the skills mentioned above. Hence, we state:

> **H2**: People with reading and writing deficits perceive proposals as more difficult to comprehend than people without these deficits.

There are several ways in which NLP can be used to help people with reading and writing deficits, such as correcting grammatical errors while writing proposals, text summarization to condense related proposals, translation into a language the reader is more familiar with to understand a proposal in more detail, or automatic text simplification to make a text in the same language as the proposal more readable. In particular, people with reading and writing disabilities could benefit from these techniques by being more certain

that they have understood a text correctly or by feeling less overwhelmed by the volume of proposals. Therefore, we postulate:

> **H3**: People with reading and writing deficits expect a higher effect/benefit through the support of NLP tasks on online deliberation platforms than people without these deficits, e.g., automatic text simplification, automatic text summarization, or machine translation.

In general, participation requires time and cognitive resources. Simplified texts could improve comprehension and thus reduce the effort required. Simplified texts also improve participation opportunities for population groups such as people with reading and writing difficulties. This creates inclusion, increases equal opportunities and improves the quality of procedures. Finally, simplified texts ensure that proposals are understood and positively received by more citizens, which supports the voting process of participatory budgeting. Therefore, we conclude:

> **H4**: The simplification of proposals has an influence on the acceptance of e-participation processes.

However, the quality of current automatic text simplification approaches in research are not ready for their usage in production. At their current state, professional post-editing by trained translators on the automatic simplified texts is mandatory (Deilen et al., 2023). A high amount of (factual) errors (Devaraj et al., 2022) and insufficient quality of the automatic simplified texts (Alva-Manchego et al., 2020) could be perceived by the readers similar as for other NLG tasks. Accordingly, it can be assumed that participants trust automatic text simplification less than manual text simplification. At the same time, automatic text simplification allows for faster, more immediate, more objective and more consistent simplification compared to simplification by a professional. Thus, participants may prefer the ability to simplify their proposals directly and anonymously, while there may be reservations about the impracticality of manual simplification, for example, in terms of paying attention to different contributions and the attention with which they are edited. Similarly, there is an increase in technology scepticism, which reduces the acceptance of automatic simplification. These factors have not yet been studied, so we postulate:

> **H5**: The influence of text simplification on the acceptance of e-participation processes is moderated by the type of simplification (`none`, `manual`, or `automatic`).

Furthermore, differences in the influence of textual simplification are to be expected depending on the role assumed. Citizens have the possibility to participate actively by writing proposals or to participate passively as readers.

On the one hand, the automatically simplified version of a text may be perceived differently by the people who wrote it. They may see the simplification as a chance for their proposals to be understood by more people with different backgrounds and thus have a wider reach. Or they may perceive the simplification as a technology-induced intrusion into their carefully crafted proposal and reject it because their own writing style has changed and they can no longer identify with their own text.

On the other hand, people who didn't write the proposal and are just reading it might welcome the simplification shown, since they have the option of reading either the original or the more readable simplified version of it. Therefore, we state as follows:

> **H6**: The influence of text simplification on the acceptance of e-participation is moderated by the role of the participants (`reader` or `author`).

4. Pre-Assumption – Readability Study

4.1. Methodology

To answer hypothesis H1, we have conducted a readability study on proposals of deliberative online participation processes. First, we have downloaded the data of online participation processes using the web crawler published in (Grawe, 2018)[1]. The resulting dataset contains overall 7,295 proposals of 11 processes (see Table 1a).

As metric to assess the readability of of the proposals, we use the German adaptation by Amstad (1978) of the Flesch Reading Ease (FRE) (Flesch, 1948) readability formula which was originally designed for English. In both languages, the score is dependent on the number of sentences, the number of words per sentence, the number of syllables per word and some language-wise constants. The formula for German FRE (Amstad, 1978) is:

$$FRE_{DE} = 180 - \frac{\#word}{\#sentences} - (58.5 * \frac{\#syllables}{\#words}) \quad (1)$$

According to this, FRE is mostly suitable for calculating the readability of documents or paragraphs. FRE is given on a scale mostly ranging

[1] The code of the web crawler is available at https://github.com/PGrawe/OnlineParticipationDatasets

between 0 and 100, where 0 is very difficult and 100 is very simple. We measured the readability with FRE (Amstad, 1978) of the Python package *textstat* [2].

However, readability metrics such as FRE have been criticized in previous work, hence, we extend our evaluation with additional linguistic metrics such as proposed in Tanprasert and Kauchak (2021). Specifically, we measure the average number of sentences per proposal, the average sentence length in words, the average word length in syllables, the average familiarity of words (measured by the average position of words in a frequency table), and the average complexity of the sentence structure (measured by the parse tree height) using SpaCy (Montani et al., 2023) and Stanza (Qi et al., 2020). For all but FRE, the lower the value, the easier it is to understand the text.

To estimate the complexity of the proposals, we can compare them with reference texts, e.g., news articles in standard German, news simplified for people with German skills following CEFR level B1, and level A2 (Council of Europe, 2020) of the APA-LHA corpus (Spring et al. (2021); see first three lines in Table 1b) or Wikipedia texts in standard German and simplified for non-native speakers of the TCDE19 corpus (Naderi et al. (2019); see last two lines in Table 1b).

4.2. Results

On average each proposal contains 4.82 ($STD = 4.08$) sentences with a maximum outlier of 9.65 sentences (see "bonn2017") and a minimum outlier of 3.92 sentences (see "raddialog-bonn"). The average sentence length is 19.42 ($STD = 7.63$) and the average word length in syllables is 1.82 ($STD = 0.22$, see Table 1a). Hence, the proposals are rather long and require discourse comprehension.

However, based on this values and following FRE, the complexity of the texts can be described as "on average" ($FRE = 42.71$, $STD = 17.91$). The process called "koeln2016" contains in comparison the most simple proposals considering a balance of sentence length, word length and number of sentences, whereas the process called "bonn2017" contains the most complex proposals.

Comparing the statistics of the proposals with the statistics of the reference corpora, the sentences of the proposals are shorter than the original news and Wikipedia texts, which is typical for user-generated content. However, the proposal sentences are on average still longer than the simplified sentences. Further, the words in the proposals are on average shorter than in the standard and simplified reference texts (except APA-LHA-A2), but as expected the words are on average more infrequent (or unfamiliar) in the proposals than in the simplified texts. In terms of structural complexity, the proposals are more nested than the simplified texts, although the proposals often do not contain complete sentences.

Overall, in line with the findings of Stodden (2021) and Bock (2015), the proposal texts appear to be difficult to understand for non-native speakers. Consequently, they may also be complex to understand for native speakers with reading difficulties. Therefore, we can confirm H1.

[2] https://github.com/textstat/textstat

name	N	S	SL↓	WL↓	FRE↑	Fam.↓	Struc.↓
badgodesberg	551	5.81	19.29	1.85	39.38	5.64	4.74
bonn2015	330	7.22	19.81	1.88	37.9	5.69	4.87
bonn2017	55	9.65	21.31	1.86	37.55	5.7	4.9
bonn2019	232	6.04	19.57	1.83	41.7	5.6	4.81
bonn2021	545	5.72	18.34	1.86	39.78	5.67	4.73
koeln2013	591	4.49	19.69	1.87	39.19	5.73	4.88
koeln2015	630	5.36	19.54	1.88	38.24	5.75	4.85
koeln2016	821	4.12	19.17	1.84	43.2	5.71	4.81
koeln2017	744	4.38	18.87	1.84	42.01	5.71	4.73
nahverkehrsplan-ulm	498	5.75	20.33	1.67	49.54	5.51	4.74
raddialog-bonn	2,298	3.92	19.57	1.78	45.82	5.65	4.7
all	7,295	4.82	19.42	1.82	42.71	5.67	4.76

(a) Overview of statistics per online participation process.

name	N	S	SL↓	WL↓	FRE↑	Fam.↓	Struc.↓
APA-LHA-OR	500	—	20.48	1.93	43.70	5.74	5.66
APA-LHA-B1	500	—	12.82	1.83	62.60	5.47	4.36
APA-LHA-A2	500	—	11.27	1.78	69.55	5.30	4.01
TCDE19-OR	250	—	25.75	2.08	28.1	5.96	6.79
TCDE19-B2	250	—	14.17	1.9	51.2	5.65	4.58

(b) Overview of statistics per reference corpus. OR = standard language, B1 and A2 = simplified language

Table 1: Overview of statistics per online participation process and reference corpora. N = number of proposals, S = number of sentences, SL = sentence length, WL= word length in syllables, FRE = Flesch Reading Ease, Fam. = Familiarity.

5. Experiment

To test the hypotheses H2 to H6, a randomized scenario-based experiment was conducted, using a 2 × 3 between-subject design with simplification style (`no` simplification or `manual` simplification or `automatic` simplification) and participant role (`reader` or `author`).

5.1. Pre-study

In order to be able to compare *no simplification* with *simplification*, we conducted a pre-study to find a complex and simplified version of the same proposal. In this pre-study, we evaluated seven different texts manually simplified into plain language (by a person trained on writing plain language) regarding their comprehensibility. All versions of the proposals can be found in Appendix A.

In order to verify the assumed readability, we asked participants to rate how comprehensible each of the variants of the following proposal is for them[3]. The scale ranges from 1 (very difficult to comprehend) to 7 (very easy to comprehend).

Overall, 21 German adults have participated in the pre-study[4]. The full results of the study are provided in Appendix A. Although version A was intended to be the most complex version, the participants have scored version B with the lowest simplicity score ($M = 4.62$, $STD = 1.56$, $N = 21$). Version G was scored with the highest simplicity score ($M = 5.86$, $STD = 1.24$, $N = 21$). Version G compared to version B shows significantly higher simplicity scores, $t(20) = 2.86, p < .01$. The effect size following Cohen (1992) is $r = .54$ and corresponds to a strong effect. Therefore, in the following study, version G will be used as the simplified version and version B as the complex, non-simplified version.

5.2. Research Design

Due to the high importance of information and communication technologies for the implementation of e-participation, information system (IS) theories such as Davis' technology acceptance model (Davis, 1985) can be used to analyze user acceptance (Naranjo-Zolotov et al., 2019). The technology acceptance model is based on the theory of planned behavior and postulates that the use of a technology is closely related to its usage intention. The usage intention, in turn, is influenced by the perceived usefulness as well as the attitude toward using, which in turn is influenced by the perceived ease of use as well as the perceived usefulness. Accordingly, we used participants' usage intention as the dependent variable to examine the influence of text simplification on the acceptance of e-participation.

We simulated a participatory budgeting system in German language using the open source platform *adhocracy+* by liquid democracy e.V.[5]. To equalize participants' experience with the platform and reduce unintended side effects, we simulate participation in the e-participation process by showing screenshots of each step of the process in an online survey.

Following, our 2 × 3 between-subject design, the participants have been grouped in six different scenarios wrt. simplification style (`no` simplification, `manual` simplification or `automatic` simplification) and participant role (`reader` or `author`) (see Table 2). All participants first read the instructions of a participatory budgeting system (almost realistic).[6] One group (called "`authors`") was simulated to write their own proposal and then read their published proposal. Another group (called "`readers`") first saw an overview of proposals and then read a published proposal of someone else (the same proposal as for the first group). For both groups, the proposals were shown in three different versions to account for the different simplification styles. Some participants saw no simplification of the proposal (see Figure 1a), and some a split screen in which the original text and the simplified text were shown side by side (see Figure 1b), following the findings by Vollenwyder et al. (2018). At the end of the simplified text, it is indicated whether the text was simplified by an employee of a translation office or by an automatic text simplification system (see "Hinweis" in Figure 1b).

We selected the text rated as most complex in the pre-study as the original (i.e., version B) and the text rated as most simple as the simplified text of the proposal (i.e., version G). This simplified text was used for both the `manual` and `automatic` simplification because we wanted to reduce the confounding variable that the different simplification transformations might cause. Further, for the same reasons, we focus on the text of one proposal and do not alter between different proposals. Participants should focus only on the translator and not on the translation, its content, or its quality.

After the simulation the participants answered a questionnaire on a 5 point Likert-scale on the following categories: demographic information, language barriers (own), acceptance of the participatory budgeting (Saura et al., 2020), perceived comprehension (Milne and Culnan, 2004), usage experience with other participatory budgeting processes (own), and NLP applications in participatory budgeting (own).

The whole study was conducted on German by Germans via the online platform Unipark[7]. The design of two different settings is visualized in Figure 1.

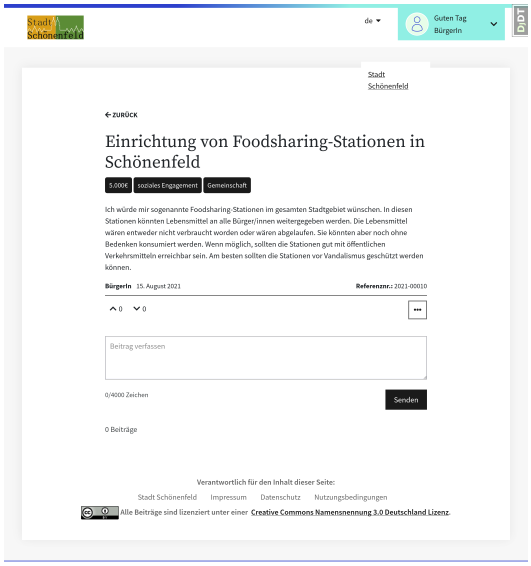
[3]The statement we asked the participants to rate was: "Bitte geben Sie an, wie verständlich die folgenden Varianten des Vorschlags für Sie sind." (engl.: "Please indicate how comprehensible the following variants of the proposal are for you.")

[4]All participants have voluntarily participated in the study.

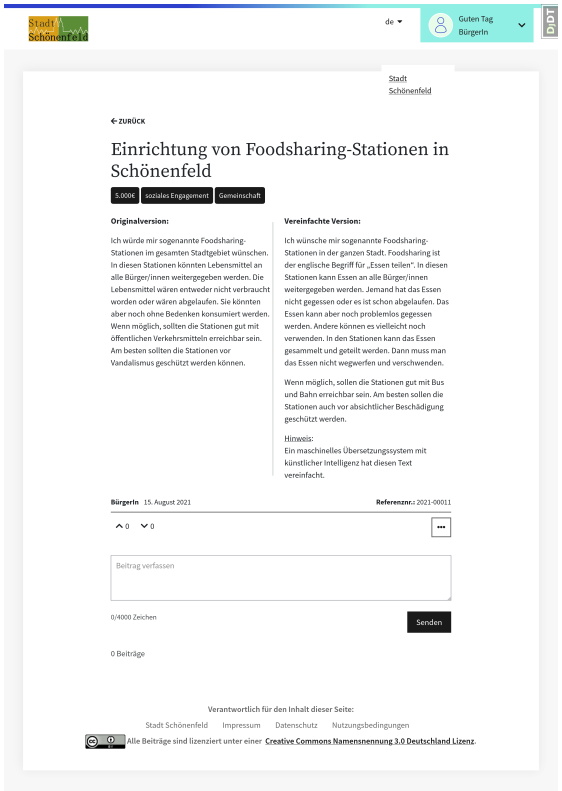
[5]`https://github.com/liqd/adhocracy-plus`

[6]The experiment is almost or nearly realistic as we provide the participants with step-by-step screenshots of the e-participation system, while they could not interact directly with the real system. To guide and control the usage of the system regarding our study design would have required an in-person experiment. The participants number in an monitored, in-person experiment would have been expected to be much lower than for the nearly realistic approach. In addition, it would be more difficult to control for confounding variables that may influence the usage intention.

[7]`https://www.unipark.com/`



(a) Only standard text.



(b) Side-by-side view of standard and plain text.

Figure 1: Two screenshots of the platform.

5.3. Results & Discussion

5.3.1. Participants

To reach the participants, study announcements were posted on various digital and analog channels such as Reddit. As an incentive, 5 Amazon gift cards were raffled. Crowdworkers were also recruited via the service provider respondi[8] to increase the sample size. 332 participants took part in the study between 2nd December 2021 and 9th March 2022.[9] After data cleaning (failed attention check), data of 276 participants could be evaluated. Overall, 60.87% of the participants are female, 38.04% male and 1.09% non-binary. 77.17% are native German speakers, 10.86% are on an expert level, 10.5% in middle level and 1.42% on beginner level. 51.09% have a Bachelor degree or higher, and 28.26% finished A-level. The average age is 38.29. 86.96% never participated in a participatory budgeting process.

Overall 25% of the participants face at least sometimes language barriers in their daily life. Furthermore, 28.99% have either low German skills, often face reading problems, are older than 65 years, or have low educational achievements. In the following we describe this group of overall 79 people as *people with reading and writing deficits* who require texts in plain language.

Scenario	Simplification	Participant
1	`none`	`reader`
2	`none`	`author`
3	`manual`	`reader`
4	`manual`	`author`
5	`automatic`	`reader`
6	`automatic`	`author`

Table 2: Overview of all groups of the experiment.

5.3.2. Attitudes Toward Readability (H2)

When combining all items regarding perceived comprehension, on average all participants rather agree that participatory budgeting proposals are overall easy to understand ($M = 3.68$, $STD = 0.66$). However, the assumption previous made (i.e., H2) that people with reading and writing deficits perceive proposals as more difficult to comprehend ($N = 79$, $M = 3.46$, $STD = 0.68$) than people without these deficits ($N = 196$, $M = 3.77$, $STD = 0.63$, $t(274) = 3.60$, $p < .01$) can be confirmed.

More specific, people with reading or writing deficits, perceive significantly more confusing terms in the proposals ($N = 79$, $M = 2.49$, $STD = 1.02$, $p < .01$, $U = 5625$, $r = .22$) and are significantly more likely to perceive the proposals as too long to be useful ($N = 79$, $M = 2.89$,

[8] https://www.respondi.com

[9] Although the study was conducted prior to the artificial intelligence hype caused by the development of ChatGPT, our results remain relevant. There is no clear indication that the use of AI tools in everyday life in Germany has changed between 2022 and 2024 (MeMo:KI, 2024).

$STD = 0.89$, $t(274) = -3.20$, $p < .01$, $r = .2$) in comparison to people without these deficits ($N = 196$, $M_{terms} = 2.00$, $STD_{terms} = 0.85$, $M_{length} = 2.51$, $STD_{length} = 0.89$). However both groups do not perceive significant comprehension differences in the structure of the proposals ($M_{without} = 3.85$, $M_{with} = 3.67$, $p > .05$). See Figure 2 for a graphical overview of the results.

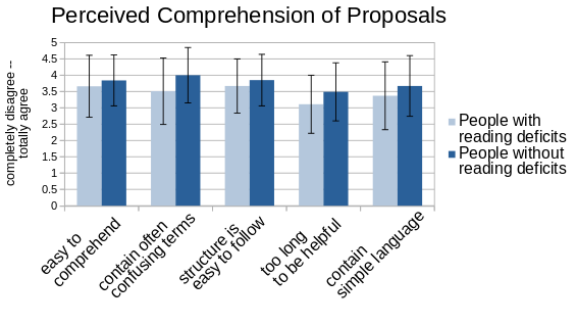


Figure 2: Mean (and standard deviation) values of people with and without reading deficits regarding perceived comprehension of proposals.

5.3.3. Attitudes Toward Text Simplification and Related NLP Tasks (H3)

Furthermore, we compared how comfortable both groups feel with automatic simplification, automatic summarization and automatic translation (all supported by artificial intelligence).

People with reading and writing deficits feel more comfortable with automatic simplification ($N = 79$; $M = 3.52$, $STD = 1.14$) than people without these deficits ($N = 196$, $M = 3.19$, $STD = 1.02$, $p < 0.05$, $t(274) = -2.30$, $r = .13$). Whereas people with reading and writing deficits feel less comfortable with automatic translation ($N = 79$, $M = 3.43$, $STD = 1.13$) than people without these deficits ($N = 196$, $M = 3.78$, $STD = 0.89$, $p < .01$, $U = 6449.5$, $r = .13$). No significant difference exist between their answers regarding automatic summarization($M_{with} = 3.46$, $STD_{with} = 1.05$, $M_{without} = 3.39$, $STD_{without} = 1.02$, $p > 0.1$). See Figure 3 for a graphical overview of the results.

In summary, both groups have same acceptance of text summarization, but people with reading and writing deficits feel more comfortable with machine simplification whereas people without these deficits feel more comfortable with machine translation.

5.3.4. Influence of Text Simplification on the Usage Intention (H4, H5, H6)

Next, we computed a two-way ANOVA to analyze the role of participation and differences of text simplification style (see Table 3). The analysis shows

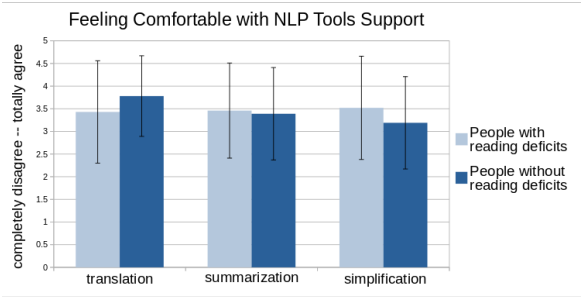


Figure 3: Mean (and standard deviation values) of people with and without reading deficits regarding support of NLP tools in e-participation processes.

Source	SS	df	MS	F	p	η_p^2
Corrected Model	1,507	5	0.301	0.500	0.776	0.009
Intercept	4037.845	1	4037.845	6696.051	0.000	0.961
TS-Type	0.578	2	0.289	0.479	0.620	0.004
Role	0.123	1	0.123	0.204	0.652	0.001
TS-Type * Role	0.763	2	0.381	0.632	0.532	0.005
Error	162.815	270	0.603			
Total	4206.200	276				
Corrected Total	164.322	275				

Table 3: ANOVA summary table for usage intention considering text simplification type (`none`, `manual`, or `automatic`) and participant role (`author` or `reader`). *SS* = Sum of Squares, *df* = degrees of freedom, *MS* = Mean Square, *F* = F value, *p* = p value, η_p^2 = Partial Eta Squared.

that there is no significant difference in the usage intention to the role of participation (`author`, `reader`; $F(1, 270) = 0.479, p > .05$; see line 4 in Table 3) or the types of text simplification (`none`, `manual`, `automatic`; $F(2, 270) = 0.204, p > .05$; see line 3 in Table 3). The interaction effect of role of participation and style of simplification on the usage intention is also not signification ($F(2, 270) = 0.632, p > .05$; see line 5 in Table 3). The usage intention therefore does not depend on the role of participation or the text simplification style. Overall, none of our hypotheses (H4, H5, and H6) could be confirmed. Since no significant effects could be detected, no post hoc-tests are necessary.

So we can transfer the results of Vollenwyder et al. (2018) also to e-participation: no unintended side-effect exists for participants without reading and writing difficulties when reading proposals in standard language and plain language side-by-side. Furthermore, we can extend these findings: the support of artificial intelligence (assuming the output of automatic text simplification systems are comparable to manual simplifications) does not negatively (or positively) influence the acceptance of e-participation, whether or not the participants wrote or only read the proposal.

6. Conclusion

In this study, we showed the need for simplification of proposals in deliberative citizen participation processes and investigated the effect of text simplification (with and without artificial intelligence support) on the acceptance of open participation processes, such as participatory budgeting. In conclusion, all results are summarized in Table 4.

H	Question	Section	Result
H1	Are proposals difficult to comprehend?	4	✓
H2	Are proposals more difficult to comprehend for people with reading difficulties than for others?	5.3.2	✓
H3	Do individuals with reading/writing difficulties expect greater benefit from simplification compared to other people?	5.3.3	✓
H4	Does the simplification of proposals have an impact on the acceptance of e-participation processes?	5.3.4	x
H5	Does the style of simplification influence e-participation acceptance?	5.3.4	x
H6	Does the role of participants influence e-participation acceptance?	5.3.4	x

Table 4: Overview of results per hypothesis.

In more detail, the results of our study show, on the one hand, that the proposals of e-participation process are difficult to comprehend (see **H1**), and even more difficult to comprehend for people with reading problems and for others (see **H2**). Further, we showed that (especially) people with reading problems welcome the assistance of NLP tools in e-participation processes, e.g., automatic text simplification, or text summarization, to overcome this issue (see **H3**).

On the other hand, we found that text simplification has no direct influence on the intention to use e-participation (**H4**). We could not find a significant effect for participants who would write or read a simplified proposal (see **H6**). Further, neither manual nor automatic simplification seem to have a effect on its usage intention (see **H5**). Although a higher level of participation cannot be achieved by text simplification, our results also show that text simplification does not have a negative influence on the intention to use e-participation. Thus, website providers do not have to fear unwanted side effects for participants without reading and writing problems. Accordingly, participants are not bothered by the simplification of their texts or those of other citizens.

Furthermore, people with reading and writing difficulties prefer text simplification for proposals in e-participation over their automatic translation or summarization. Hence, adding automatic text simplification systems to e-participation processes can be a meaningful contribution by making it easier for many people to understand complex proposals.

7. Limitations & Future Work

Unfortunately, the number of participants with reading and writing deficits in our experiment was comparatively small, which may be due to similar reasons as the low participation rate in online participation processes, e.g., this target group is difficult to reach. Due to the small number of participants with reading and writing deficits in this study, no highly reliable or causal statements can be made, but they do provide initial indications of the importance of text simplification in the context of e-participation.

Furthermore, the idiosyncrasies of the proposal text might have an effect of the results, in future work, the study could be repeated with alternating simplification of different proposals.

In future work, we would like to conduct a qualitative field experiment with people with reading and writing deficits which might evoke less barriers as a direct contact person could be present to give further explanations similar as described in the study by Säuberli et al. (2024). Further, in this experiment, the usage of real generated simplifications of a text simplification system could be tested, which was currently not possible, because current state-of-the-art German text simplification systems do not support user-generated texts or online participation proposals. Based on the results of our study, further investigation into text simplification for proposals in deliberation processes, or for user-generated texts in general, would be worthwhile.

8. Bibliographical References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Toni Amstad. 1978. *Wie verständlich sind unsere Zeitungen?* Studenten-Schreib-Service. PhD Thesis.

Miguel Arana-Catania, Felix-Anselm Van Lier, Rob Procter, Nataliya Tkachenko, Yulan He, Arkaitz Zubiaga, and Maria Liakata. 2021. Citizen participation and machine learning for a better democracy. *Digit. Gov.: Res. Pract.*, 2(3).

Hadi Asghari, Freya Hewett, and Theresa Züger. 2023. On the prevalence of leichte sprache on the german web. In *15th ACM Web Science Conference 2023*, WebSci '23, New York, NY, USA. Association for Computing Machinery.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In

Proceedings of the Sixth International Joint Conference on Natural Language Processing, pages 356–364, Nagoya, Japan. Asian Federation of Natural Language Processing.

Bettina M. Bock. 2015. Barrierefreie Kommunikation als Voraussetzung und Mittel für die Partizipation benachteiligter Gruppen: Ein (polito-)linguistischer Blick auf Probleme und Potenziale von "Leichter" und "einfacher Sprache". *Linguistik Online*, 73(4).

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical Error Correction: A Survey of the State of the Art. *Computational Linguistics*, 49(3):643–701.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing, Strasbourg, France. Available at www.coe.int/lang-cefr.

Fred D Davis. 1985. *A technology acceptance model for empirically testing new end-user information systems: Theory and results*. Ph.D. thesis, Massachusetts Institute of Technology.

Silvana Deilen, Sergio Hernández Garrido, Ekaterina Lapshinova-Koltunski, and Christiane Maaß. 2023. Using ChatGPT as a CAT tool in easy language translation. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 1–10, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345, Dublin, Ireland. Association for Computational Linguistics.

Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

European Telecommunications Standards Institute. 2021. Accessibility requirements for ICT products and services - EN 301 549 (V3.2.1). https://www.etsi.org/deliver/etsi_en/301500_301599/301549/03.02.01_60/en_301549v030201p.pdf.

Delia Ferri and Silvia Favalli. 2018. Web accessibility for people with disabilities in the european union: Paving the road to social inclusion. *Societies*, 8(2):1–19.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Dennis Friess, Katharina Esau, and Christiane Eilders. 2017. How emotions, humor and narratives interact with traditional characteristics of deliberation online. In *Proceedings of the 67th ICA Annual Conference*, volume Panel Paper, San Diego, USA.

Silke Gutermuth. 2019. *Leichte Sprache für alle? Eine zielgruppenorientierte Rezeptionsstudie zu Leichter und Einfacher Sprache*. Frank & Timme, Berlin, Germany.

Loni Hagen, Özlem Uzuner, Christopher Kotfila, Teresa M. Harrison, and Dan Lamanna. 2015. Understanding citizens' direct policy suggestions to the federal government: A natural language processing and topic modeling approach. In *Proceedings of the 48th Hawaii International Conference on System Sciences*, pages 2134–2143, Kauai, HI, USA. IEEE.

Steven L. Johnson, Hani Safadi, and Samer Faraj. 2015. The emergence of online community leadership. *Information Systems Research*, 26(1):165–187.

Joyce Karreman, Thea Van Der Geest, and Esmee Buursink. 2007. Accessible website content guidelines for users with intellectual disabilities. *Journal of Applied Research in Intellectual Disabilities*, 20(6):510–518.

Matthias Liebeck, Katharina Esau, and Stefan Conrad. 2016. What to do with an airport? mining arguments in the German online participation project tempelhofer feld. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 144–153, Berlin, Germany. Association for Computational Linguistics.

MeMo:KI. 2024. Dashboard of the Public Opinion Monitor Artificial Intelligence [Meinungsmonitor Künstliche Intelligenz]. Available at https://www.cais-research.de/forschung/memoki/.

George R. Milne and Mary J. Culnan. 2004. Strategies for reducing online privacy risks: Why consumers read (or don't read) online privacy notices. *Journal of Interactive Marketing*, 18(3):15–29.

Mijail Naranjo Zolotov, Tiago Oliveira, and Sven Casteleyn. 2018. E-participation adoption models research in the last 17 years: A weight and meta-analytical review. *Computers in Human Behavior*, 81:350–365.

Mijail Naranjo-Zolotov, Tiago Oliveira, and Sven Casteleyn. 2019. Citizens' intention to use and

recommend e-participation. *Inf. technol. people*, 32(2):364–386.

Eleni Panopoulou, Efthimios Tambouris, and Konstantinos Tarabanis. 2014. Success factors in designing eparticipation initiatives. *Information and Organization*, 24:195–213.

Peter Parycek, Michael Sachs, Florian Sedy, and Judith Schossboeck. 2014. Evaluation of an e-participation project: Lessons learned and success factors from a cross-cultural perspective. In *Proceedings of the 2014 International Conference on Electronic Participation (ePart 2014)*, pages 128–140, Berlin & Heidelberg, Germany. Springer.

Julia Romberg and Tobias Escher. 2023. Making sense of citizens' input through artificial intelligence: A review of methods for computational text analysis to support the evaluation of contributions in public participation. *Digit. Gov.: Res. Pract.* Just Accepted.

Jose Ramon Saura, Pedro Palos-Sanchez, and Felix Velicia-Martin. 2020. What drives volunteers to accept a digital platform that supports ngo projects? *Frontiers in Psychology*, 11:1–14.

Armin Schäfer and Harald Schoen. 2013. Mehr Demokratie, aber nur für wenige? Der Zielkonflikt zwischen mehr Beteiligung und politischer Gleichheit. *Leviathan*, 41(1):94–120.

Sven Schmutz, Andreas Sonderegger, and Juergen Sauer. 2019. Easy-to-read language in disability-friendly web sites: Effects on nondisabled users. *Applied Ergonomics*, 74:97–106.

Anthony Simonofski, Monique Snoeck, Benoit Vanderose, Joep Crompvoets, and Naji Habra. 2017. Reexamining e-participation: systematic literature review on citizen participation in e-government service delivery. In *Proceedings of the 23rd Americas Conference on Information Systems*, pages 1–10, Boston, USA. Association for Information Systems (AIS).

Regina Stodden. 2021. Accessibility and comprehensibility of user-generated content: Challenges and chances for easy-to-understand languages. In *Proceedings of the 3rd Swiss Conference on Barrier-free Communication (BfC 2020)*, pages 151–161, ZHAW Zürcher Hochschule für Angewandte Wissenschaften. ZHAW Zurich University of Applied Sciences.

Andreas Säuberli, Franz Holzknecht, Patrick Haller, Silvana Deilen, Laura Schiffl, Silvia Hansen-Schirra, and Sarah Ebling. 2024. Digital comprehensibility assessment of simplified texts among persons with intellectual disabilities.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14, Online. Association for Computational Linguistics.

Beat Vollenwyder, Andrea Schneider, Eva Krueger, Florian Brühlmann, Klaus Opwis, and Elisa D. Mekler. 2018. How to use plain and easy-to-read language for a positive user experience on websites. In *Computers Helping People with Special Needs*, pages 514–522, Cham. Springer International Publishing.

Hui Yang, Jamie Callan, and Stuart Shulman. 2006. Next steps in near-duplicate detection for erulemaking. In *Proceedings of the 2006 International Conference on Digital Government Research*, dg.o '06, page 239–248, San Diego, California, USA. Digital Government Society of North America.

Robert Zepic, Marcus Dapp, and Helmut Krcmar. 2017. Participatory budgeting without participants: Identifying barriers on accessibility and usage of german participatory budgeting. In *Proceedings og the 2017 Conference for E-Democracy and Open Government (CeDEM)*, pages 26–35, Krems, Austria. IEEE.

9. Language Resource References

Philipp Grawe. 2018. *Analyse von Online-Partizipationsverfahren : Themenextraktion, Visualisierung und Interaktion*. Master Thesis, Heinrich-Heine-University.

Ines Montani, Matthew Honnibal, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. explosion/spaCy.

Babak Naderi, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller. 2019. Subjective assessment of text complexity: A dataset for german language. *CoRR*, abs/1904.07733:1–4.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Nicolas Spring, Annette Rios, and Sarah Ebling. 2021. Exploring German multi-level text simplification. In *Proceedings of the International*

Conference on Recent Advances in Natural Language Processing (RANLP 2021), pages 1339–1349, Held Online. INCOMA Ltd.

A. Pre-study

Version	Text	Mean	Std
A	Ich würde mir wünschen, dass im gesamten Stadtgebiet Foodsharing-Stationen aufgestellt werden, in denen nicht verbrauchte oder abgelaufene Lebensmittel, die noch ohne Bedenken konsumiert werden können, an alle Bürger/innen weitergegeben werden könnten. Die Stationen sollten nach Möglichkeit gut mit öffentlichen Verkehrsmitteln erreicht und vor Vandalismus geschützt werden können.	5.48	1.6
B	Ich würde mir sogenannte Foodsharing-Stationen im gesamten Stadtgebiet wünschen. In diesen Stationen könnten Lebensmittel an alle Bürger/innen weitergegeben werden. Die Lebensmittel wären entweder nicht verbraucht worden oder wären abgelaufen. Sie könnten aber noch ohne Bedenken konsumiert werden. Wenn möglich, sollten die Stationen gut mit öffentlichen Verkehrsmitteln erreichbar sein. Am besten sollten die Stationen vor Vandalismus geschützt werden können.	4.62	1.56
C	Ich wünsche mir sogenannte Foodsharing-Stationen im gesamten Stadtgebiet. In diesen Stationen können Lebensmittel an alle Bürger/innen weitergegeben werden. Die Lebensmittel sind entweder nicht verbraucht worden oder sind abgelaufen. Sie könnten aber noch ohne Bedenken konsumiert werden. Andere können das Essen vielleicht noch verwenden. Wenn möglich, sollen die Stationen gut mit öffentlichen Verkehrsmitteln erreichbar sein. Am besten sollen die Stationen auch vor Vandalismus geschützt werden.	5.52	1.29
D	Ich wünsche mir sogenannte Foodsharing-Stationen in der ganzen Stadt. In diesen Stationen kann Essen an alle Bürger/innen weitergegeben werden. Das Essen ist entweder nicht gegessen worden oder es ist abgelaufen. Das Essen kann aber noch problemlos gegessen werden. Andere können es vielleicht noch verwenden. Wenn möglich, sollen die Stationen gut mit Bus und Bahn erreichbar sein. Am besten sollen die Stationen auch vor absichtlicher Beschädigung geschützt werden.	5.38	1.2
E	Ich wünsche mir sogenannte Foodsharing-Stationen in der ganzen Stadt. In diesen Stationen kann Essen an alle Bürger/innen weitergegeben werden. Jemand hat das Essen nicht gegessen oder es ist schon abgelaufen. Das Essen kann aber noch problemlos gegessen werden. Andere können es vielleicht noch verwenden. Wenn möglich, sollen die Stationen gut mit Bus und Bahn erreichbar sein. Am besten sollen die Stationen auch vor absichtlicher Beschädigung geschützt werden.	5.19	1.36
F	Ich wünsche mir sogenannte Foodsharing-Stationen in der ganzen Stadt. In diesen Stationen kann Essen an alle Bürger/innen weitergegeben werden. Jemand hat das Essen nicht gegessen oder es ist schon abgelaufen. Das Essen kann aber noch problemlos gegessen werden. Andere können es vielleicht noch verwenden. In den Stationen kann das Essen gesammelt und geteilt werden. Dann muss man das Essen nicht wegwerfen und verschwenden. Wenn möglich, sollen die Stationen gut mit Bus und Bahn erreichbar sein. Am besten sollen die Stationen auch vor absichtlicher Beschädigung geschützt werden.	5.33	1.2
G	Ich wünsche mir sogenannte Foodsharing-Stationen in der ganzen Stadt. Foodsharing ist der englische Begriff für „Essen teilen“. In diesen Stationen kann Essen an alle Bürger/innen weitergegeben werden. Jemand hat das Essen nicht gegessen oder es ist schon abgelaufen. Das Essen kann aber noch problemlos gegessen werden. Andere können es vielleicht noch verwenden. In den Stationen kann das Essen gesammelt und geteilt werden. Dann muss man das Essen nicht wegwerfen und verschwenden. Wenn möglich, sollen die Stationen gut mit Bus und Bahn erreichbar sein. Am besten sollen die Stationen auch vor absichtlicher Beschädigung geschützt werden.	5.86	1.24
B (translated)	I would like to see so-called food sharing stations throughout the city. At these stations, food could be passed on to all citizens. The food would either not have been consumed or would have expired. However, they could still be consumed without hesitation. If possible, the stations should be easily accessible by public transportation. Ideally, the stations should be able to be protected from vandalism.		
G (translated)	I would like to see so-called foodsharing stations all over the city. Foodsharing is the English term for “sharing food”. At these stations, food can be passed on to all citizens. Someone has not eaten the food or it has already expired. But the food can still be eaten without any problem. Others may still be able to use it. In the stations the food can be collected and shared. Then there is no need to throw the food away and waste it. If possible, the stations should be easily accessible by bus and train. Preferably, the stations should also be protected from intentional damage.		

Table 5: Simplified variants of the proposal and its simplicity scores of the pre-study.

Pitfalls of Conversational LLMs on News Debiasing

Ipek Baris Schlicht[1,2], **Defne Altiok**[1], **Maryanne Taouk**[3], **Lucie Flek**[4,5,6]
[1]Deutsche Welle, Bonn/Berlin, Germany
[2]Universitat Politecnica de Valencia, Spain
[3]ABC News, Australia
[4]Conversational AI and Social Analytics (CAISA) Lab
[5]Bonn-Aachen International Center for Information Technology (b-it), University of Bonn
[6]The Lamarr Institute for Machine Learning and Artificial Intelligence

Abstract

This paper addresses debiasing in news editing and evaluates the effectiveness of conversational Large Language Models in this task. We designed an evaluation checklist tailored to news editors' perspectives, obtained generated texts from three popular conversational models using a subset of a publicly available dataset in media bias, and evaluated the texts according to the designed checklist. Furthermore, we examined the models as evaluator for checking the quality of debiased model outputs. Our findings indicate that none of the LLMs are perfect in debiasing. Notably, some models, including ChatGPT, introduced unnecessary changes that may impact the author's style and create misinformation. Lastly, we show that the models do not perform as proficiently as domain experts in evaluating the quality of debiased outputs.

Keywords: News Bias Correction, LLMs, Human Evaluation, Automatic Evaluation

1. Introduction

Biased news articles have the potential to significantly shape public opinion and discourse on various issues. Thus, professional news editors identify bias text spans in news articles before they are published. This task is particularly challenging, especially when editorial teams face constraints such as time-pressure and a lack of human resources.

Large Language Models (LLMs) have demonstrated outstanding performance even in the absence of labeled data, through zero-shot prompting. In many tasks, LLMs have surpassed the performance of the supervised models and have even employed as writing assistance (Shi et al., 2022; Zhang et al., 2023). In addition, conversational LLMs such as ChatGPT (OpenAI, 2022) and GPT4 (OpenAI, 2023) are user-friendly, making them accessible to non-technical experts like journalists who can use them without coding knowledge to aid in their tasks. As a result, many media companies have already begun experimenting ChatGPT for various journalistic tasks (Beckett, 2023). Limited studies have explored debiasing through text generation with conversational LLMs for the tasks such as hate speech (Plaza-del arco et al., 2023) and toxicity detection(Morabito et al., 2023). These studies explored zero-shot prompting with conversational LLMs. To the best of our knowledge, conversational LLMs have not been explored for news debiasing.

Standard evaluation metrics (Min et al., 2023) such as ROUGE require a reference text for measuring generated text quality and lacks explanatory evaluation. Morabito et al. (2023) established an evaluation protocol for automatically assessing LLMs' consistency in debiasing for toxicity detection by using Perspective API(per) as the evaluator. However, this protocol is limited to bias reduction and may not be suitable for the news domain. In the context of news bias, bias encompasses both overt bias such derogatory terms within text and latent biases that shape the language and framing of news stories (Recasens et al., 2013). As shown in Figure 1, news texts deemed biased may not contain toxicity but wording/phrasing could introduce bias. Hence, tools such as Perspective API could fail to quantify bias reduction. Furthermore, the debiased text might produce misinformation by changing context and factuality and altering the author's writing style. Therefore, there is a need for

President Donald Trump gave states and local governments the right to reject refugees, but instead of saying no, most state and local officials have blindsided the administration by opting in, according to two former officials familiar with the matter.
President Donald Trump allowed states and local governments the option to refuse refugees. However, according to two former officials familiar with the matter, most state and local officials have chosen to accept refugees.

Figure 1: Biased text where the usage "blind-sided" introduces bias by conveying a strong negative opinion about the actions of state and local officials and its GPT4 debiased version which doesn't contain toxicity according to Perspective API. Debiasing changed the facts and the context (factually incorrect statement highlighted in red, original version in blue).

The First Workshop on Language-driven Deliberation Technology (DELITE2024), pages 33–38

ID	Concept	Question
C1	**Correcting Bias**	**Does the model produce unbiased text? Grade 1-3** *The text is unbiased. (3)* *The text is partially biased. (2)* *The text is highly biased. (1)*
C2	**Preserving Information**	**Does the model change textual facts? Grade 1-3** *The text facts are still present. (3)* *Some facts are missing. (2)* *Facts are completely missing. (1)*
C3	**Preserving Context**	**Does the model change the meaning of text? Grade 1-3** *The meaning of the text is preserved.(3)* *The meaning of the text is partially preserved. (2)* *The meaning of the text is completely changed.(1)*
C4	**Preserving Language Fluency**	**Does the model produce grammatically correct text? Grade 1-3** *The text is grammatically correct. (3)* *The text has few grammar issues. (2)* *The text has many grammar issues. (1)*
C5	**Preserving Author's Style**	**Does the model harm the author's creativity? Grade 1-3** *No, the model did all necessary changes without harming author creativity. (3)* *The model corrected some of the texts that might hurt the creativity. (2)* *The model did unnecessary changes, and changed the text style. (1)*

Table 1: News editorial criteria for checking quality of debiasing.

evaluation criteria discerned editorial perspectives.

To address these issues, we investigate the following research questions (RQs): (1) How well do conversational LLMs perform debiasing in the context of the news domain according to editorial criteria? (2) Can conversational LLMs also serve as an evaluation tool for assessing the editorial quality of debiased articles?

Given the need for a domain-specific evaluation to assess the quality of conversational LLMs in news debiasing, we propose a set of evaluation criteria tailored to news editors. Since there is no publicly available news dataset for debiasing, we obtained text generations on a subset of the publicly available bias classification dataset using three popular conversational LLMs and a fine-tuned T5 (Raffel et al., 2020). Expert news editors from international media organizations ranked the models' outputs based on the editorial criteria. Additionally, we compared model outputs with expert assessments when the models were used as evaluation tools to check the quality of debiasing. Our results showed that despite conversational LLMs' proficiency in bias reduction, they sometimes generate misinformation and alter writing styles. Moreover, they can not assess debiased outputs as the experts do [1].

2. Related Works

The studies on media bias have primarily focused on two aspects: identifying biased text spans (Spinde et al., 2021; Hamborg, 2020; Lei et al., 2022) and detecting political bias in news articles (Chen et al., 2020) or media outlets (Baly et al., 2020). Only a few studies proposed methods for mitigating bias through article generation using transformer models. Among these studies, the earliest work by Pryzant et al. (2020) used BERT to identify subjective content and update the hidden layers of the model to generate unbiased text from Wikipedia. Lee et al. (2022) applied a summarization method on articles from various political leanings to neutralize news.

Plaza-del arco et al. (2023) and Morabito et al. (2023) explored the potential of zero-shot prompting with LLMs, respectively for hate speech detection and reducing toxicity in user comments. Additionally, Morabito et al. (2023) established an evaluation protocol for evaluating consistency of LLMs on debiasing in the context of toxicity detection. The authors used Perspective API as the evaluator tool which provides toxicity scores for comment moderation. However, the protocol is limited to only to bias reduction. Furthermore, is not applicable within the news domain as news articles may not exhibit a toxic tone, yet they can still contain biases favoring certain groups, which need to be addressed before publication. In our work, we design evaluation criteria taking into account journalistic perspective to measure quality of debiased sentences.

Recently, researchers have explored LLMs as evaluators for assessing the quality of text generation in various applications (Gao et al., 2024; Min et al., 2023) as an alternative solution to costly expert assessments. Motivated by this, we evaluate

[1]The code and the data are at https://bit.ly/3vGphbw

the conversational LLMs models as evaluators for assessing the quality of debiased sentences based on the journalistic criteria and compare them with our expert evaluation.

3. Methodology

We investigated three conversational LLMs for debiasing news sentences and paragraphs. Given sentences or paragraphs containing bias types such as epistemological, framing and demographic bias (Pryzant et al., 2020; Spinde et al., 2021; Recasens et al., 2013), the goal of the task was to generate an unbiased version of those sentences. The outcome of the sentences should be unbiased but other criteria should also be considered as important for news editors, such as preserving factuality, news' message, and not harming the authors' creativity, along with grammar changes.

3.1. News Editorial Criteria

As prior evaluation metrics are limited to news debiasing, we propose news editorial criteria. The editorial criteria were created during the implementation of BiasBlocker, which is a prototype AI-based news editor. [2].

The BiasBlocker team comprises experienced news editors and technologists from Deutsche Welle, ABC News and ARIJ. Since bias is a broad concept, to establish a common ground on the bias definitions and the corrections, the editors in the team created a codebook on bias types [3] and guidelines for debiasing based on the prior studies (Pryzant et al., 2020; Spinde et al., 2021; Recasens et al., 2013) and UN Guidelines [4]. Hence, the bias types we focus on are primarily framing, epistemological, and demographic bias.

We applied a pilot study on bias correction by using ChatGPT with the editors [5]. The editors spotted the issues and refined the expectations for the news editor. As outlined in Table 1, we distilled these expectations into five criteria for assessing the quality of models in the context of debiasing for news editing.

Essentially, the editors expected the model to effectively remove any text spans that introduce bias into the content. However, they also had the expectation that this must refrain from adding new facts or removing vital information, as this could produce misinformation. Furthermore, the model must ensure that the meaning of the text remains intact. The debiased text must also be grammatically correct. Lastly, especially for those articles of opinion pieces or analyses, the model must respect and preserve the author's writing style and creativity. Otherwise, the model could discourage less experienced authors and harm the communication of the news message.

Evaluation Dataset. Wiki Neutrality Corpus (WNC) (Pryzant et al., 2020) is the only publicly available dataset that contains biased samples and their debiased version by Wikipedia editors. Given that our research objective was to assess the LLMs in correcting bias within texts authored by news authors, WNC samples were not suitable for the evaluations. Therefore, we preferred the BABE dataset (Spinde et al., 2021) as the test set. BABE consists of sentences from news articles published by US publishers with different political leanings. Experienced media experts annotated the dataset; the dataset samples were labeled as biased or unbiased. The authors of the dataset provided two subsets. We chose the one annotated with more experts and randomly selected 50 biased sentences from this subset for the evaluations.

3.2. Debiasing Models

Baseline. As the baseline, we used the large version of T5 (Raffel et al., 2020). T5 is an encoder-decoder transformer that is pre-trained on a cleaned Common Crawl collection, incorporating a mixture of supervised tasks through multi-task learning. To adopt T5-large for the debiasing task, we used WNC as the training dataset. Given our constraints with computational resources, we fine-tuned the model using LoRA adaptation (Hu et al., 2022).

Conversational LLMs. We evaluated the popular conversational LLMs: ChatGPT (OpenAI, 2022) and GPT4 (OpenAI, 2023) from OpenAI, and Llama2-70b-chat (Touvron et al., 2023) which is an open source, popular alternative to ChatGPT. The models were adopted for conversational tasks using reinforcement learning with human feedback. In this way, individuals without technical expertise could easily interact with the models, making them suitable for integration into news organizations.

ChatGPT and Llama 2 are Autoregressive Language Models trained on large corpora from multiple sources from the web, with the objective of predicting the next word based on the preceding context. GPT4 is the advanced version of ChatGPT, capable of handling multi-modal input. While our task focused on textual input, we included GPT4 in our evaluations, because human evaluators preferred GPT4 outputs from various tasks over ChatGPT (OpenAI, 2023). We used prompts, which are shown in Table 2, for each of the conversational.

[2] https://bit.ly/4aJttWD
[3] https://bit.ly/49qcnvZ
[4] https://bit.ly/3PRks67
[5] https://bit.ly/43OMCnQ

Debiasing Prompt

Transform the following biased sentence into an unbiased sentence from a news article by removing any subjective language or discriminatory undertones without changing its semantic meaning:

Biased Sentence:

{{sentence}}

Unbiased Sentence:

Evaluator Prompt (shortened)

The input sentence from a news article is biased, it uses subjective language or discriminatory undertones. The other sentence was debiased by a language model. Your task is to compare two sentences based on the following journalistic criteria. For each question in the checklist, select your response from {1, 2, 3}.

The checklist is as follows:
1- Does the model produce unbiased text?
- If the text is unbiased, return 3
...

Do not explain your decisions.

Biased Sentence:

{{sentence}}

Model Output:

{{model_output}}

Checklist Answers:

Figure 2: Prompts for debiasing and evaluation. The full version of the evaluator prompt can be found at our source code.

4. Results

Although BABE contains the biased text spans along with the labels, the dataset does not have the corrected versions of the biased texts. Therefore, we could not directly apply the evaluation criteria to the samples. For this reason, two expert news editors from the team, as described in § 3.1, conducted the human evaluations voluntarily. Due to resource constraints, we split the models' results into two parts for both evaluators. Each part contains the results from each model. One editor ranked the samples which they were responsible for, by using a 3-likert scale. During the ranking evaluation, the editor marked the samples they were unsure about, made notes and applied fact-checking to address the C2 and C3. The other editor reviewed the ranked samples while checking the notes, marked samples and the fact-checked ones. The editors regularly engaged in discussions to reach a consensus on disagreements and uncertain cases. In total, we obtained 200 evaluations from the experts. Table 2 presents the frequency of ratings per criterion.

ID	Grade	T5	Llama2	ChatGPT	GPT4
C1	1	0.26	0.08	0.02	0
	2	0.40	0.06	0.14	0.38
	3	0.34	**0.86**	0.84	0.62
C2	1	0.1	0.4	0.26	0.2
	2	0.12	0.44	0.36	0.56
	3	**0.78**	0.16	0.38	0.24
C3	1	0.12	0.34	0.20	0.06
	2	0.06	0.4	0.48	0.68
	3	**0.82**	0.26	0.32	0.26
C4	1	0.34	0.1	0.02	0
	2	0.2	0	0	0.12
	3	0.46	0.9	**0.98**	0.88
C5	1	0.14	0.44	0.42	0.42
	2	0.08	0.46	0.56	0.5
	3	**0.78**	0.1	0.02	0.08

Table 2: The conversational LLMs are significantly better than the baseline at correcting bias and providing grammatically correct outputs (Student's T-test, p-value at 0.05), they have issues on preserving information, context and author's style.

RQ1: Debiasing Performance of the Conversational Models. The conversational LLMs proved better than the baseline for debiasing. Surprisingly, Llama 2 demonstrated comparative results even though ChatGPT has been known to outperform others in various tasks (Touvron et al., 2023). The researchers of Llama 2's training regime - that the factual sources were prioritized in training samples - might account for its competitive performance in this task. The conversational LLMs also exhibited more grammatical correctness than the baseline. Nevertheless, some LLMs changed phrases they considered biased, while others removed words or sentiments that could be considered confrontational or impolite, but are not actually biased towards any particular group. For instance, GPT4 changed 'When carrying a firearm, you have the ultimate power of force in your control' to 'When carrying a firearm, you have a significant level of potential force at your disposal'.

The conversational LLMs performed worse than the baseline model in preserving information and context. These models introduced unnecessary amendments to the generated texts. In some cases, even created hallucinations. This issue is not unique to this study and has been reported in related studies, especially in the case of ChatGPT being used for various tasks (Bang et al., 2023).Additionally, the news editors observed that Llama 2

introduced additional information not present in the input text, albeit factually accurate. For example, in a text mentioning 'Wilkens', the model replaced 'Wilkens' with 'Judge Wilkens'. The model may have memorized such information from its training dataset. This behavior by conversational models might harm the author's style.

ID	Llama2	ChatGPT	GPT4
C1	0.0666	-0.0489	0.1109
C2	-0.0145	0.0285	0.0018
C3	0.1971	0.0280	0.0263
C4	0.0597	0	0
C5	-0.0022	0.0454	-0.0413

Table 3: The disagreement between the conversational tools as an evaluator and the expert evaluation is high, according to Cohen's Kappa.

RQ2: Conversational LLMs as Evaluator: We obtained rankings from the conversational LLMs and compared them with the expert rankings. As shown in Table 3, there is a high disagreement between the models and the expert evaluations. Additionally, we observed that the models rated the criteria, such as preserving factuality, grammar, with the highest score. In contrast, the ratings by the experts for these criteria were low.

5. Conclusion

Through the editorial criteria, we showed that none of the conversational LLMs are perfect, even though they are good at debiasing. Specifically, they failed to preserve vital information and context, often leading to hallucinations. Employing these tools in a fully automatic editor can be dangerous, as they can create misinformation.

Memorization also surfaces as an important aspect of LLM behavior. For this reason, to ensure a fair evaluation of debiasing tasks across news articles from different periods, Media bias researchers need to create benchmark datasets containing samples from time periods that is not covered within the training data of LLMs.

The assessments by the models are not close to those by the experts. We plan to increase the size of our annotations and the number of annotators to build a benchmark dataset for a fine-grained analysis of the models' issues. We then investigate advanced methods for automating the evaluation criteria and incorporating them to adapt the models.

Ethical Considerations and Limitations

In this study, we assessed the efficiency of conversational LLMs in debiasing news articles, focusing solely on English samples from US Media. As a result, the generalizability of our conclusions to other languages and to media in other countries may be limited.

The dataset employed in this research paper is derived from publicly accessible sources and is peer-reviewed. During the evaluation process, we refrained from disclosing the identities of the article publishers to the participating news editors, thereby preventing any potential influence on their evaluations.

Acknowledgement

This research was partially funded by the JournalismAI Fellowship Programme 2024 of PolisLSE and vera.ai, which is co-financed by the European Union, Horizon Europe programme, Grant Agreement No 101070093. We also thank Kevin Nguyen, Saja Mortada, Khalid Waleed for their support.

References

Perspective api. https://perspectiveapi.com/. Accessed: 2024-02-15.

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020. What was written vs. who read it: News media profiling using text analysis and social media context. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3364–3374.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023.

Charlie Beckett. 2023. How newsrooms around the world use ai: a journalismai 2023 global survey. https://blogs.lse.ac.uk/polis/2023/06/26/how-newsrooms-around-the-world-use-ai-a-journalismai-2023-global-survey/.

Wei-Fan Chen, Khalid Al Khatib, Henning Wachsmuth, and Benno Stein. 2020. Analyzing political bias and unfairness in news articles at different levels of granularity. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 149–154, Online. Association for Computational Linguistics.

Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024. Llm-based NLG evaluation: Current status and challenges. *CoRR*, abs/2402.01383.

Felix Hamborg. 2020. Media bias, the social sciences, and nlp: automating frame analyses to identify bias by word choice and labeling. In *Proceedings of the 58th annual meeting of the association for computational linguistics: student research workshop*, pages 79–87.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net.

Nayeon Lee, Yejin Bang, Tiezheng Yu, Andrea Madotto, and Pascale Fung. 2022. Neus: Neutral multi-news summarization for mitigating framing bias. In *NAACL-HLT*, pages 3131–3148. Association for Computational Linguistics.

Yuanyuan Lei, Ruihong Huang, Lu Wang, and Nick Beauchamp. 2022. Sentence-level media bias analysis informed by discourse structures. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10040–10050, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Robert Morabito, Jad Kabbara, and Ali Emami. 2023. Debiasing should be good and bad: Measuring the consistency of debiasing techniques in language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4581–4597, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2022. Openai: Introducing chatgpt. https://openai.com/blog/chatgpt.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *ACL (1)*, pages 1650–1659. The Association for Computer Linguistics.

Shuming Shi, Enbo Zhao, Duyu Tang, Yan Wang, Piji Li, Wei Bi, Haiyun Jiang, Guoping Huang, Leyang Cui, Xinting Huang, Cong Zhou, Yong Dai, and Dongyang Ma. 2022. Effidit: Your AI writing assistant. *CoRR*, abs/2208.01815.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. Neural media bias detection using distant supervision with BABE - bias annotations by experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yue Zhang, Leyang Cui, Deng Cai, Xinting Huang, Tao Fang, and Wei Bi. 2023. Multi-task instruction tuning of llama for specific scenarios: A preliminary study on writing assistance. *CoRR*, abs/2305.13225.

Integrating conflict prevention tools into deliberative democracy online platforms

Sara Greco, Chiara Jermini-Martinez Soria
Università della Svizzera italiana USI
Via Buffi 13 – 6900 lugano (Switzerland)
sara.greco@usi.ch, chiara.jermini@usi.ch

Abstract

This paper presents a set of preliminary guidelines for conflict prevention developed within the EU-funded research project ORBIS ("Augmenting participation, co-creation, trust and transparency in Deliberative Democracy at all scales") whose goal is developing online platforms that enable citizens to enhance their participation in democratic processes, through open discussions around important political topics. Based on previous research on communication and argumentation in conflict resolution discourse and on the empirical analysis of discussions around deliberative democracy topics, this paper highlights recurrent interpersonal communication problems that might occur in group discussions around complex topics and that, if not handled well, can lead to conflicts; and introduces a first proposal for solutions to help, both through technology and with the assistance of human moderations, participants in such discussions to avoid the development and the escalation of conflicts.

Keywords: conflict resolution, deliberative democracy, argumentation, dialogue

1. Purpose of the guidelines

This paper presents preliminary results concerning of conflict detection, prevention and resolution tools based on discourse and argumentation into online **platforms for deliberative democracy**. It is not uncommon, in fact, for conflicts between individuals or sub-groups to arise in group discussions around important topics that impact people's lives. While having different positions and ideas around a specific issue is normal and it can potentially be a good thing (from disagreements may arise new ideas or solutions for problems), when people start to 'fight' with the intent to prevail over the other there are usually negative consequences (see Greco 2020). Namely, the interaction may shift from an opportunity to have a fruitful discussion around the exploration of realistic solutions to common problems or ways to reach common goals, to an argument in which the main goal is to 'win' by discrediting the other (see Greco 2020). At the theoretical level, this paper is based on research on argumentation in conflict resolution and professional dispute mediation of interpersonal conflicts (e.g. Aakhus, 2003; Greco, 2011; van Bijnen and Greco, 2018; Greco, Cigada & Jermini-Martinez Soria, 2024). On the basis of this research, we have first analyzed an empirical dataset of discussions regarding deliberative democracy organizations, identifying needs for conflict prevention. Subsequently, we have turned to a design perspective, suggesting possible tools for conflict prevention to be integrated in deliberative democracy online platforms as based on needs emerged from the organizations' talk. While this paper presents preliminary results of an ongoing project and draft solutions, it also shows a possible methodology to design conflict prevention tools in online platform as based on the requests explicitly or implicitly emerging from participants to discussion., .

2. Methodology

The methodology used for designing these guidelines is based on a qualitative approach, using a cycle that goes from deductive, to inductive, and cycles back to deductive qualitative research. The process was inspired by Bingham (2023) and adapted to the specific goals of these guidelines. In particular, the methodology designed for these guidelines includes three steps: 1) We derived *inductively*, from previous research on conflict resolution, possible sources of conflict and misunderstanding. This research step is based on our team's existing research and analysis of a dataset of conflict resolution interactions concerning interpersonal conflicts on several topics (around 180K words); 2) We compared the possible sources for conflict with the existing five datasets collected within the project. While the type of discussion is different, we can assume that some discursive elements of conflict escalation are present across different fields and geographical areas (e.g. Greco, 2011). On the basis of this comparative evaluation, we identified the main emerging needs for conflict detection, prevention and resolution. The annotation was jointly developed by the two authors; the first round of annotation was done by author 1; all problematic cases were then discussed by both authors to find joint interpretations and verify reliability of the annotation. We also identified needs emerging from the datasets that were not present in our original list of sources of conflict; thus, we included an *inductive* dimension in the research; 3) Going back to our research on conflict resolution, we connected each emerging need to possible discursive conflict resolution tools, reflecting on how these could be integrated into platforms for deliberative democracy. For step 2, the researchers have analyzed five anonymized cases constituting "the ORBIS dataset"; three of them that have been collected in the project to elicit organizations' needs for tools for online deliberation platforms (and three *Building Blocks for Democracy* events organized by project partners. All

The First Workshop on Language-driven Deliberation Technology (DELITE2024), pages 39–44

these five discussions have some elements in common: first, they were all conducted in English, which guarantees a common basis for the linguistic-argumentative analysis of the interaction. Second, they all included meta-reflection by participants regarding conflict resolution and what they would like to see implemented in the platforms for deliberative democracy in order to improve the quality of the interactions . Before proceeding, it is necessary to specify they did not find direct indication of existing conflict during the recorded discussions but rather "emerging needs" for possible conflict resolution prevention tools. With the term *need* for conflict prevention mechanisms, we indicate clues that show potential or real problems that the parties have identified in their experience of deliberative democracy, such as moments of *impasse* (see Aakhus, 2003) in the discussion, problematic situations that lead to potential frustration or conflictual behavior. The goal of the online platforms for deliberative democracy that are being developed within the ORBIS project is both to ensure good quality interactions among users of the platforms, in order to guarantee respectful and inclusive exchanges, and to come up with concrete proposals to foster positive social changes: in this sense, the identification of "needs" is then followed by a design reflection on what features would be desirable to have in online platforms in order to increase the possibilities to have such a positive discussion environment and to reach the goals of deliberative democracy. The process of annotation was done using Microsoft Excel to annotate data from the datasets listed above: we reported in an Excel file excerpts in which emerging needs for conflict prevention, management and resolution were present, classifying the type of need according to the categories described in Section 3.

3. Needs for conflict prevention

In this section, we will discuss the main needs for conflict preventions that we have identified in the ORBIS datasets, namely:

1. Silence and lack of participation
2. Issue management and need to shift from the problem to the options
3. Lack of common ground
4. Presence of dysphoric emotions
5. Who should participate in the discussion?

3.1 Silence and lack of participation

One of the issues that participants describe and that generate a need for conflict prevention is the difficulty of including a wide range of 'voices' in the discussions, as it frequently happens that it is always the same - restricted - group of people that engages in the deliberative democracy talk. This category regards a basic problem in communication, namely participants' active presence. It is a category we derived inductively from the analysis of the ORBIS dataset, In larger group discussions, the problem of silence and lack of participation is an important indicator of possible conflict. The presence of silent members who do not express their opinions may be problematic for different reasons: if the people who talk are somehow representatives of the same 'group' (e.g. gender, ethnicity,...) the opinions and concerns of minorities can go unnoticed; if people do not overtly express their opinions it is difficult to guess whether this means they agree with what has been already said or not; people with specific and valuable knowledge about a relevant aspect of a context are not involved in the discussion and so on. Silence could therefore be an indication of the presence of a "cold conflict" (Greco, 2020), in which people cultivate resentment or disillusionment.

3.2 Issue management and need to shift from the problem to the options

Since the issues addressed by participants to democratic discussions are mostly complex political ones, oftentimes it becomes difficult to structure a discussion in a way that is functional to the emergence of concrete and feasible proposals for solutions to address said issues. For example, each participant might insist on focusing only on one specific element (e.g. time, when to do something) of an issue without considering other crucial aspects. For moderators, it is difficult to balance the freedom of participants to express themselves and the need for efficacy.

As it is known from previous research, in conflict resolution typically, after the analysis of the dispute, it is important to move on to discussing possible *options* for its resolution (Greco, 2011, p. 75); this is part of the issue management that conflict mediators operate to create a discussion space conducive to a productive resolution of differences of opinion. After having listened to different points of views, it's difficult to enter a phase where concrete proposals that everyone agrees with are to be made. To be able to do so, it is fundamental to make sure that all parties' interests are duly taken into account, otherwise the discussion will not move forward as people might perceive that 'giving up' one's own idea corresponds to the risk of not having their interests and needs met (see Greco, 2011).

3.3 Lack of common ground

When many different people discuss together about a complex topic, it is inevitable that they have different levels of knowledge about each aspect of it. This might create misunderstandings and confusion among participants, as one might not understand well what someone else is saying or why. This issue is partially linked to "issue management" (3.2), in the sense that to be able to have a fruitful discussion about a complex topic, it is fundamental to establish a solid common ground *first* - otherwise people risk addressing the topic only with their partial and not overlapping knowledge in mind, which might create frustration and confusion in other participants.

3.3.1 Identification of interests underlying positions

Very often, when involved in disagreements or debates, people will clearly express their positions, i.e. their point of view regarding an *issue* - for example whether they are in favor or against a political decision, and they will argue to support their thesis. However, through a deep discussion based on active listening and questioning, also their needs might emerge. We typically hold a position because we think that is 'the solution', i.e. what will have our needs met - but that might not be the only answer. It is not easy for people to spontaneously focus on their interests leaving aside their positions because "people's egos become identified with their positions" (Fisher, Ury and Patton, 1991, p. 11).

3.3.2 Conflicting frames and getting to know 'the others', their points of view

To have a fruitful and open discussion, it is fundamental to consider one's interlocutors as 'worthy' and value their ideas (see the concept of "exploratory talk", Mercer and Littleton, 2013). For this to be possible, each person needs to believe that it is possible for different points of views to coexist and to be equally valid. This does not mean that everyone will agree with the others, but that decentrating and listening to the others' point of view is a prerequisite for having a respectful and open discussion. Participants need to understand the 'frame' of the others to be able to further discuss and possibly come up with ideas that will be acceptable to all (Shmueli, 2008; Mercuri, 2023).

3.4 Dysphoric emotions

From the analysis of the dataset, it is clear that the topics related to democratic participation addressed by the use cases participants are usually topics perceived as personally important (e.g. unemployment) and therefore addressing them might provoke difficult emotions in participants that, if not handled well, might even trigger conflicts. As Jones (2001) explains, "emotion results from a perception that something personally important is at stake" (Jones, 2001, p. 94), and, therefore "the triggering events that 'cause' conflicts are, by definition, events that elicit emotion" (Jones, 2001, p. 90).

3.5 WHO should participate/not participate in the discussion?

From our inductive analysis of the data, it emerged that, in order to be able to find solutions that can really be implemented, it is crucial that all the people who have a say/a decision-making role in the matter are actually involved in the discussion process.

4. Tools that can be used in the project's deliberative democracy platforms

In this section, we move on to a perspective of design to make hypotheses about tools that can be potentially integrated into online platforms for deliberative democracy (Table 1) based on the needs emerging from the analysis of the ORBIS dataset and described in Sect. 3. The shift from the analysis of conflict prevention needs to the design proposal requires interdisciplinary collaboration, which is currently ongoing with technical partners of the ORBIS project; therefore, what we propose in this paper is a preliminary set of hypotheses that still need to be implemented and tested. At the moment, we draw the hypothesis that these tools may be based partially on automatic alerts and generative AI, while partially they may require the presence of a human moderator. While the tasks performed automatically could also be attributed to a human moderator, since the goal is to create online tools that function in the best way possible automatically, it makes sense to limit the intervention of human moderators to perform tasks that cannot be performed by the technology. Moreover, it is important to point out that a moderator is *not* a professional mediator, as s/he has not received the same specifical professional training – however some of the communicative techniques employed by professional dispute mediators are also useful in the context of moderation. All tools proposed (both automatic alerts and human moderations) have been drawn from existing and well-established literature on discourse and argumentation in conflict resolution and dispute mediation (see section 1). The discussion about implementation is still in progress in ORBIS.

Emerging needs	**Possible tools (automatic alerts)**	**Possible tools (human moderation)**
Silence and participation	**Questions** to engage participants who are silent (after 5 minutes, ask question to silent participants: do you want to say something?)	**Alerts** can be sent to a human moderator after 5 minutes a person has been silent. The moderator will decide what to do.
Need for issue management (includes: need to shift from the problem to the options)	**Automatic summaries** of what has been said are made visible but does not interfere with the progress of the discussion. **Regular reminders** of the main issue of the discussion are shown.	**Reformulations**: if negative terms are present in the summaries, the moderator can decide to change the terms from negative to a more neutral formulation. For example, a sentence such as "The other party does not understand the situation" can be reformulated with "I understand that there are different perspectives on the understanding of this situation".

		Often, reformulations can be done by moving from verbs or adjectives to nouns (**nominalization**), for example: - Participant A is wrong → "There is a different perspective on the issue between you and A" - "These people create problems" → There is a problem that you all need to discuss. A human moderator can also be trained to decide when it is the moment to **shift from the problem to the options for its resolution**, for example: - "Now you have acknowledged that you have a problem of management of participations of young people. How do you think you can resolve it? Do you have suggestions?"
Lack of common ground (includes: confusion between interests and positions Conflicting frames and need to know the other)	**Automatic prompts asking for meta-reflection** to the involved participants. These prompts must be visible to each party individually but not to the others: - Is there anything you want to ask to the other participants to better understand their positions? - Is there anything you want to add about your position and why you are proposing it? All this requires adding a space for **Clarification questions to the other participants**	**Open, explorative and non-accusatory questions** on the parties' profound reasons behind their positions. For example: - Do you feel you could explain your position enough? Do you want to add something more? - How would you describe the reasons why you have this position? Is there any aspect you want to share? - Is there anything you want to know about someone else's positions and why they hold it?
Presence of dysphoric emotions	–	**Open, explorative and non-accusatory questions** on emotions, including dysphoric emotions. When there is a negative emotion, it should not be covered but the parties need to be given space to explain what are the reasons behind the emotions. This needs to be done by a human moderator because it is a delicate option, which risks to create escalation. A possible tool to introduce these questions is the "**I hear you say" intervention** (van Bijnen and Greco, 2018), in which the human moderator can say for example "I hear you say that you are concerned, can you explain

		more?". This mitigates the question and offers an opportunity to the speaker to explain the reasons behind emotions, opening up a space for argumentation that includes personal worries. Dysphoric emotions are often related to feelings of guilt and resentment. The literature shows that a potential useful tool in these cases is **reframing from individual to system** (Putnam, 2004) or **from individual to species** (Jermini, 2021). This means reminding to the parties that the problem does not necessarily concern them only but it can concern other people. This helps removing blame and feelings of guilt. Possible formulation of this reframing are: - "All the citizens who are interested in their cities have this sort of problem" - "I feel that also other participants to this discussion sooner or later had to face this problem"
Who should participate in the discussion?	Adding a **List of further participants that would help solving the issue** could help to see whether	Human moderators can decide how to organize the following sections taking into account the list proposed by the participants.
	important stakeholders are missing. The system might automatically ask participants to think about this issue, for example by asking: "can we make a decision on this topic or someone else need to be involved?"	

Table 1: Emerging conflict prevention needs and possible tools to be integrated into the platforms

Working together with technical experts within the EU-funded ORBIS project, the next step will be to select the tools that can more easily be implemented in online platforms for deliberative democracy and to have use cases participants test the platforms including these integrations, in order to see if they are effective to prevent conflict escalation and how they can further be improved.

5. References

Aakhus, M. (2003). Neither naïve nor critical reconstruction: Dispute mediators, impasse, and the design of argumentation. *Argumentation*, 17(3): 265-290. https://doi.org/10.1023/A:1025112227381

Bingham, A. J. (2023). From data management to actionable findings: A five-phase process of qualitative data analysis. *International Journal of Qualitative Methods*, *22*. https://doi.org/10.1177/16094069231183620Fisher, R., Ury, W., and Patton, B. (1991, 2nd ed.). G*etting to yes. Negotiating agreement without giving in.* Penguin Books, New York, 2nd edition.

Greco, S. (2011). *Argumentation in dispute mediation: A reasonable way to handle conflict*. John Benjamins, Amsterdam.

Greco, S. (2020). *Dal conflitto al dialogo: Un approccio comunicativo alla mediazione*. Maggioli Editore, Santarcangelo di Romagna.

Greco, S., Cigada, S., and Jermini-Martinez Soria, C. (2024). The naming of emotions in dispute mediators' strategic manoeuvring. A case study using a French language corpus. *Text & Talk,* 44(1). https://doi.org/10.1515/text-2021-0044

Jermini-Martinez Soria, C. (2021). *Reframing as an argumentative competence in dispute mediation.* PhD dissertation, Università della

Svizzera italiana (USI), Switzerland. https://doc.rero.ch/record/330915

Jones, T. S. (2001). Emotional communication in conflict: Essence and impact. In W. F. Eadie and P. E. Nelson (Eds.), *The language of conflict and resolution*. Thousand Oaks: Sage, pp. 81-104.

Mercer, N. and Littleton, K. (2013). *Interthinking: Putting talk to work*. Routledge, London.

Mercuri, C. (2023). Characterization frames constructing endoxa in activists' discourse about the public controversy surrounding fashion sustainability. *Topoi*, 42(2): 635-650. https://doi.org/10.1007/s11245-023-09909-2

Shmueli, D. F. (2008). Framing in geographical analysis of environmental conflicts: Theory, methodology and three case studies. *Geoforum*, 39(6): 2048-2061. https://doi.org/10.1016/j.geoforum.2008.08.006

Van Bijnen, E., and Greco, S. (2018). Divide to unite: Making disagreement explicit in dispute mediation. *Journal of Argumentation in Context*, 7(3): 285-315. https://doi.org/10.1075/jaic.17032.bij

A Hybrid Human-AI Approach for Argument Map Creation From Transcripts

Lucas Anastasiou, Anna De Liddo
Knowledge Media Institute, The Open University
{lucas.anastasiou, anna.deliddo}@open.ac.uk

Abstract

In order to overcome challenges of traditional deliberation approaches that often silo information exchange between synchronous and asynchronous modes therefore hindering effective deliberation, we present a hybrid framework combining Large Language Models (LLMs) and human-in-the-loop curation to generate argument maps from deliberation transcripts. This approach aims to enhance the efficiency and quality of the generated argument maps, promote transparency, and connect the asynchronous and synchronous deliberation modes. Finally, we outline a realistic deliberation scenario where this process can be successfully integrated.

1. Introduction

Deliberation processes are important mechanisms for collaborative decision-making, fostering informed choices across a wide array of domains (Vaculín et al., 2013; Owen, 2015). Traditionally, these processes occurred through either synchronous (in-person or real-time online) discussions or asynchronous (such as online discussion forums) (Wright and Street, 2007). However, the distinction to synchronous and asynchronous consists of a siloed approach to deliberation that creates barriers to information exchange, development of shared understanding and subsequently consensus building and other elements that consist of effective deliberation (Friess and Eilders, 2015).

Recent advancements in Natural Language Processing (NLP) and particularly in Large Language Models (LLMs) have created promising paths to structure and synthesise information such as unstructured dialogue, i.e. free-flowing conversation (e.g. transcripts of meetings, online chat conversations) or semi-structured data (e.g. interviews, XML documents, and others) (Naveed et al., 2023; Serban et al., 2016). They possess the potential to generate structured discourse data (e.g. argument graphs or key points) (Chen et al., 2023). This may be the unblocker to overcome some of the challenges associated with traditional deliberative processes. Nonetheless, despite their impressive performance, LLMs are not without limitations; they are still susceptible to misinterpretation (Turpin et al., 2024), hallucinations (Ye et al., 2023), inaccuracies (Guo et al., 2023), therefore making them unreliable to be used in sensitive applications (like public deliberation that has significant impact in decision making).

To address these shortcomings, we propose an approach involving a human-in-the-loop (HITL) model ((Zanzotto, 2019)) to curate and filter LLM-generated outputs before integrating them into asynchronous debate platforms. This paper explores the potential of this hybrid framework to bridge the gap between synchronous and asynchronous deliberation modes, promoting accountability, transparency and more accurate and informed decision making.

2. Related work

2.1. Imperfect AI

Despite rapid advancements in the field, AI systems remain imperfect and likely will continue to be for the foreseeable future. Errors can arise from biases embedded in training data, limitations of the algorithms themselves, or unpredictable real-world inputs (Suresh and Guttag, 2019; Mehrabi et al., 2021). Furthermore, the "black-box" nature of many deep learning models hampers explainability, obscuring the logic behind potentially erroneous outputs (Samek et al., 2019). This persisting imperfection highlights the critical need for human oversight and intervention, especially in high-stakes domains. Recent work by Bussone et al. (2015) demonstrates how faulty AI-generated explanations can even worsen the situation, leading to unwarranted trust and potentially harmful decisions. Therefore, in contexts where safety, accuracy, and fairness are paramount, human-in-the-loop approaches remain essential for mitigating risks, ensuring ethical outcomes, and ultimately fostering responsible AI integration (Lee et al., 2020; Leslie, 2019).

2.2. Human-AI collaboration

Human-AI is focusing on the effective integration of human intelligence with the power of AI. Such collaboration holds the potential to surpass the limitations of either humans or AI working alone (Wilson and Daugherty, 2018; Passi and Vorvoreanu, 2022).

The First Workshop on Language-driven Deliberation Technology (DELITE2024), pages 45–51

To achieve various levels of collaboration, workflows such as human-in-the-loop (HITL), where AI provides assistance with humans retaining decision authority, and human-on-the-loop (HOTL), focusing on constant human oversight, have seen extensive exploration (Liu et al., 2014). Additionally, recent studies advocate for a human-in-command approach (Wesche and Sonderegger, 2019; Bostrom and Yudkowsky, 2018) stressing the necessity of maintaining ultimate human control in critical applications.

2.3. Argument mining using LLMs

Argument mining, the task of identifying and extracting argumentative structures from text (Cabrio and Villata, 2018; Lawrence and Reed, 2020), has seen significant advancements with Large Language Models (LLMs), such as OpenAI's GPT-3 (Brown et al., 2020), GPT-4 (OpenAI, 2023), Google's Gemini (Team et al., 2023), Anthropic's Claude (Azzollini and Pomponio, 2019) and others. LLMs' ability to understand and generate complex language enables more nuanced argument extraction (Kashefi et al., 2023), offering the potential to improve argument component identification, relationship classification, and even argumentative summarisation (Reimers et al., 2019; Lauscher et al., 2022; Elaraby and Litman, 2022). This opens up opportunities for automated analysis of large-scale debates, supporting decision-making, and facilitating critical thinking. However, challenges remain. LLMs can conflate correlation with causation, leading to the identification of spurious arguments (Jin et al., 2023). Additionally, biases inherent in the LLM's training data can propagate into argument identification (Acerbi and Stubbersfield, 2023). Despite these limitations, LLM-based argument mining holds significant promise for understanding and structuring complex discourse.

3. Proposed Method

We propose a curated method for argument map creation from conversational data (specifically from transcripts of informal or formal meetings) that prioritises both accuracy and automation, combining the capabilities of computational tools and the critical reasoning by humans. This hybrid Human-AI approach involves:

- *Initial AI Processing*: We utilise LLM prompting to mine arguments (identify argumentative components) from the transcript of conversations. The transcripts consist of written records of what was said in a meeting, speech, interview or any other spoken event; in our case we use video captions (.srt files) that is easily accessible (though not ideal as there is no speaker identification). We build the argument map using the simplified IBIS model ((Kunz and Rittel, 1970)), i.e. organising arguments into positions and pro (supporting) or con (opposing) arguments. An illustrative method for extracting arguments from textual transcripts using Large Language Models (LLMs) to the Issue-Based Information System (IBIS) argumentation scheme is shown in Prompt 1. Note that to facilitate transparency and provenance, we emphasize the inclusion of original transcript snippets alongside generated arguments.

- *Human Annotation and Curation*: At this stage the generated argument map is presented to a human curator where they annotate each argument node across several evaluation dimensions inspired by Argument Mining evaluation frameworks (e.g. Sofi et al. (2022)) such as *Groundedness* (Levonian et al. (2023) - whether the argument generated is based on the input text), *Context Relevance* (whether it draws from the surrounding text - it relates to the connected argument) and others. Such annotation process can be logged using modern software such as trulens[1]. Human curators are enabled to confirm the inclusion of each argument node, edit the content of it or change the connection links to each. To facilitate this process we use several visual assistance aids that we explain further in Section 3.2. The curated versions of the argument maps are later used to as ground truth examples to finetune the LLM used in the initial AI processing stage.

- *Semantically connect and merge with other argument maps*: At this stage we proceed to import into the curated argument map into an established database of argument maps/debates. We identify similar arguments by comparing the semantic similarity of the argument nodes (using e.g. argueBERT (Behrendt and Harmeling, 2021)). We proceed to merge the similar arguments following again a curation workflow (asking humans to select whether to combine the two arguments by generating via LLM a summary of the two or just denote explicitly the similarity of both but keep separated)

- *Key-Point analysis and summarisation*: Upon creating the final argument map, we proceed to create a summarised view, i.e. automatically extracting the core arguments or essential messages from the collection of arguments (using key point analysis (Bar-Haim et al., 2020)).

[1] https://www.trulens.org/

Prompt 1 Extract key positions and argument from transcript

Below is a transcript from a debate in the european parliament:
–
{{ TRANSCRIPT TEXT FROM SRT FILE }}
–
What are the main positions and arguments for and against given in the above? Provide those in a bulleted list like:
- Position N: <position_text>
– Arguments supporting Position N (pro arguments):
— <argument text N.p.i>
– Arguments against Position N (con arguments):
— <argument text N.c.j>
Do not include supporting or opposing arguments if they do not exist. Make sure you include only arguments or positions that appear in the given text. To make sure that this is the case, on each argument or position include the timestamp that this is mentioned in the given text

3.1. Example

We present here the output of the application of such prompting in a sample taken from "Economic Dialogue with Christine Lagarde"[2] in the European Parliament in Figure 1. Our analysis revealed significant variation in the outputs generated by the different models used. As expected, GPT3.5, exhibits the weakest performance, producing a comparatively simplistic representation of the arguments presented in the example transcript. Interestingly, the outputs from the two more proficient models (GPT4 and Gemini Advanced) displayed distinct characteristics. It is noteworthy that Gemini Advanced deviated from the instructed format and fully omitted any counter-arguments (con arguments) from its representation.

3.2. Curation workflow and interface

The output of the initial AI processing, while demonstrating promising accuracy, cannot guarantee perfect results. Therefore, we propose a following human-in-the-loop curation workflow that incorporates several design elements:

1. *Confidence Indicators*: We display confidence levels derived from the probability of each generated token, presented visually as a highlight with red color in case of low logit probability. As an extension, we propose calculating the semantic distance between the generated argument component and the original transcript snippet, serving as a proxy for confidence, and display this confidence score on each generated argument node.

2. *Alternative Outputs*: The workflow provides a selection of alternative outputs for curators to choose from, enhancing decision-making and refining the extracted arguments. Alternative outputs can come from invocation of different models (as shown in our example, outputs significantly vary according to the used model), use of different prompts or chunking strategies.

3. *Editable Argument Tree with Drag-and-Drop Functionality*: Users can directly manipulate the generated argument tree. Transcript chunks or individual arguments can be repositioned within the tree, allowing for intuitive restructuring and refinement.

4. *Approval and Confidence Assessment*: Each argument node offers either a simple approval/rejection mechanism (checkmark/cross x mark) or a granular confidence slider (0-100%). In case of multiple annotators, we can follow a simple majority agreement to ensure reliabel results.

Crucially, for points 3 and 4, the system retains a historical log of actions (approvals, changes, dismissals). This historical record is a crucial element for ensuring transparency, accountability, and ultimately, trust in the quality of the final curated argument structure - as it allows the reconstruction of decision-making processes and providing rich explanations for the final curated argument structure.

3.3. Evaluation design

For the systematic evaluation of the workflow described in Section 3.2 we consider the following hypotheses:
H1: Different LLM models impact the quality of generated argument maps
H2: Human-in-the-loop curated argument maps exhibit an improvement in quality compared to argument maps generated solely by automated methods

Regarding dataset, we shall focus on transcripts for argument mining tasks that are deliberation contextualised (e.g. M-Arg Mestre et al. (2021)). Both H1 and H2 can be tested by asking domain experts (human annotators) to evaluate the quality and coherence of argument extractions from each LLM candidate and with or without human curation or/and with the use of computational metrics (accuracy, F1) against the reference dataset. This can be

[2] https://multimedia.europarl.europa.eu/en/video/economic-dialogue-with-christine-lagarde-ecb-president-opening-statements-by-irene-tinagli-sd-it-chair-of-econ-and-introductory-remark-by-christine-lagarde-president-of-the-european-central-bank-ecb$_I252842$

(i) GPT3.5

Position 1: Concern about inflation levels
– Arguments supporting Position 1 (pro arguments):
— Inflation fell to 2.4% in November 2023, reaching its lowest level since July 2021 (Chapter 1)
— Inflation increased again to 2.9% in December 2023 (Chapter 1)
— The recent increase was mainly due to upward base effects for the cost of energy (Chapter 2)
— Inflation levels are considerably lower compared to the average annual inflation in 2023 (Chapter 2)
— ECB is determined to ensure that inflation returns to its 2% medium-term target (Chapter 2)
Position 2: Impact of ECB's monetary policy on inflation
– Arguments supporting Position 2 (pro arguments):
— Tight financing conditions are dampening demand and helping to push down inflation (Chapter 3)
— ECB's interest rate increases are being transmitted forcefully into financing conditions (Chapter 3)
— ECB's forward-looking wage structure signals strong wage pressures, but agreements suggest some levelling (Chapter 11)
— Wage pressure for 2024 hinges particularly on ongoing and upcoming negotiations (Chapter 11)
— Contribution of unit profits to domestic price pressures continued to decline, suggesting that wage increases are buffered by profit margins (Chapter 12)

(ii) GPT4

Position 1: Inflation monitoring and response by the ECB
Arguments supporting Position 1:
— Inflation fell to 2.4% in November and increased to 2.9% in December, showing fluctuating but controlled levels (00:00:02.000).
— After ten rate hikes, the ECB left the key interest rate unchanged for the third consecutive time, showing a determined stance to ensure inflation returns to the 2% medium-term target (00:00:48.000).
— Tight financing conditions are helping to push down inflation, indicating the effectiveness of ECB policies (00:01:45.000).
Position 2: The euro's significance and performance over 25 years
Arguments supporting Position 2:
— The euro area has grown from 11 to 20 countries, demonstrating the currency's expanding influence and integration within Europe (00:02:16.000).
— Support for the euro among citizens is close to record levels, reflecting its acceptance and perceived benefits (00:02:16.000).
— The ECB's monetary policy has been instrumental in navigating financial and COVID crises, underscoring the central bank's role in maintaining economic stability (00:03:16.000).

(iii) Gemini Advanced

Position 1: The euro has been a success story, delivering simplicity, stability, and sovereignty to its member states.
Arguments supporting Position 1 (pro arguments):
Simplicity: Eliminates currency exchange hassle within the eurozone, facilitating trade and travel. (Chapters 14, 16)
Stability: Relatively low average inflation (2.1% since 1999) and anchored long-term inflation expectations signal ECB success in maintaining price stability. (Chapter 16)
Sovereignty: As a major global currency, the euro insulates eurozone economies from international currency fluctuations. (Chapter 17)
Position 2: The European Central Bank (ECB) is committed to ensuring the euro's continued success.
Arguments supporting Position 2 (pro arguments):
Mandate focus: The ECB has responded forcefully to inflation surges, demonstrating a commitment to price stability. (Chapter 18)
Climate adaptation: The ECB is incorporating climate change considerations into monetary policy operations. (Chapter 19)
Digital innovation: The ECB's ongoing work on the digital euro shows a willingness to adapt to technological shifts. (Chapter 20)

Figure 1: Example of argument map generations from transcript of Economic Dialogue with Christine Lagarde in European Parliament. Variations of output using (i) GPT3.5, (ii) GPT4 and (iii) Gemini Advanced. The presented is a fraction of the generated output due to space constraints

achieved through surveys in an A/B experimental setup.

4. Use case

Further to the systematic evaluation, we envision to incorporate the above described method into a deliberation scenario where a policy organisation utilises the LISTEN-REFRAME-ACT (L-R-A)[3] method to broader citizen and expert engagement on public policy issues. The L-R-A method is a structured apporach to public deliberation that: The *LISTEN* phase emphasizes on deep understanding of the diverse perspectives surrounding an issue. In the *REFRAME* phase, based on the insights from the LISTEN phase, the participants focus on reframing the issue collaboratively, developing more inclusive, evidence-based narratives and exploring potential solutions. In the last *ACT* phase, the reframed understanding and ideas are transformed into actionable proposals.

Traditionally all of the above phases are carried out in physical settings. The proposed transcript-to-argument-graph conversion method offers a powerful solution. It enables importing *LISTENING* phase insights directly into an online deliberation platform. By systematically analyzing transcripts, extracting key arguments, and incorporating LLM-assisted refinement, this method enables the successful transition from unstructured discussions to argumentative structure discussion. The generated argument maps can be used to seed further focused online discussions, providing a grounded starting point for the *REFRAME* and *ACT* phases. This integration ensures that the valuable insights from the *LISTENING* phase are effectively carried forward into the online deliberation, enhancing the process's richness and inclusivity.

5. Conclusions

This paper has presented an approach for argument map creation from transcript text that offers a synergistic approach, combining the efficiency of computational automation with the depth of human critical thinking, therefore getting results superior to either in isolation. Our method empowers untrained users to effectively construct argument maps, addressing a known challenge highlighted

[3] https://www.linkedin.com/pulse/future4citizens-barcelona-european-capital-democracy-xxgge/

in prior research (e.g. Le et al., 2013). Crucially, our approach maintains human control throughout the process, ensuring transparency and accountability in the resulting argument map. This fosters trust between users and the generated outcomes. Moreover, this method has the potential to facilitate the fluid exchange from synchronous to asynchronous deliberation modes. Future development could explore the integration of chain-of-thought (Wei et al., 2022) or tree of thoughts (Yao et al., 2024) reasoning for improving the performance of the AI pre-processing and also mitigate dependence on prompt engineering. Importantly, while the method seeks to mitigate individual subjectivity through majority agreement, human annotation of what consists argument or position remains inherently susceptible to personal perspectives. Finally, the scope of this work did not include the addition of arguments into an existing knowledge base, leaving room for exploration into how the approach can support the evolution of established argument maps. Future work will focus on implementing and extending this approach in real large-scale deliberation scenarios.

6. Acknowledgements

This research was funded in collaboration by UKRI under the UK Government's Horizon Europe Guarantee scheme (Reference Number: 10048874) and by the European Commission under the Horizon Europe Programme, in the context of the ORBIS Project (GA: 101094765) on "Augmenting participation, co-creation, trust and transparency in Deliberative Democracy at all scales".

7. Bibliographical References

Alberto Acerbi and Joseph M Stubbersfield. 2023. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44):e2313790120.

Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine*, 35(4):105–120.

Antonio Azzollini and Alessio Pomponio. 2019. Positive energy static solutions for the chern-simons-schrödinger system under a large-distance fall-off requirement on the gauge potentials.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From arguments to key points: Towards automatic argument summarization. *arXiv preprint arXiv:2005.01619*.

Maike Behrendt and Stefan Harmeling. 2021. Arguebert: How to improve bert embeddings for measuring the similarity of arguments. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 28–36.

Nick Bostrom and Eliezer Yudkowsky. 2018. The ethics of artificial intelligence. In *Artificial intelligence safety and security*, pages 57–69. Chapman and Hall/CRC.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*, pages 160–169. IEEE.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: A data-driven analysis. In *IJCAI*, volume 18, pages 5427–5433.

Guizhen Chen, Liying Cheng, Luu Anh Tuan, and Lidong Bing. 2023. Exploring the potential of large language models in computational argumentation. *arXiv preprint arXiv:2311.09022*.

Mohamed Elaraby and Diane Litman. 2022. Arglegalsumm: Improving abstractive summarization of legal documents with argument mining. *arXiv preprint arXiv:2209.01650*.

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

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab,

and Bernhard Schölkopf. 2023. Can large language models infer causation from correlation? *arXiv preprint arXiv:2306.05836*.

Omid Kashefi, Sophia Chan, and Swapna Somasundaran. 2023. Argument detection in student essays under resource constraints. In *Proceedings of the 10th Workshop on Argument Mining*, pages 64–75.

W. Kunz and H.W.J. Rittel. 1970. *Issues as Elements of Information Systems*. Number no. 131 in California. University. Center for Planning and Development Research. Working paper, no. 131. Institute of Urban and Regional Development, University of California.

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

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

Nguyen-Thinh Le, Sabine Niebuhr, David Drexler, and Niels Pinkwart. 2013. Scaffolding computer supported argumentation processes through mini map based interaction techniques. In *Design, User Experience, and Usability. Web, Mobile, and Product Design: Second International Conference, DUXU 2013, Held as Part of HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part IV 2*, pages 533–542. Springer.

Min Kyung Lee, Nina Grgic-Hlaca, Michael Carl Tschantz, Reuben Binns, Adrian Weller, Michelle Carney, and Kori Inkpen. 2020. Human-centered approaches to fair and responsible ai. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*.

David Leslie. 2019. Understanding artificial intelligence ethics and safety. *ArXiv*, abs/1906.05684.

Zachary Levonian, Chenglu Li, Wangda Zhu, Anoushka Gade, Owen Henkel, Millie-Ellen Postle, and Wanli Xing. 2023. Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference. *arXiv preprint arXiv:2310.03184*.

Juan Liu, Aaron Wilson, and David Gunning. 2014. Workflow-based human-in-the-loop data analytics. In *Proceedings of the 2014 Workshop on Human Centered Big Data Research*, pages 49–52.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Rafael Mestre, Razvan Milicin, Stuart E. Middleton, Matt Ryan, Jiatong Zhu, and Timothy J. Norman. 2021. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *Proceedings of the 8th Workshop on Argument Mining*, pages 78–88, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

OpenAI. 2023. Gpt-4 technical report.

Daniel Owen. 2015. Collaborative decision making. *Decision Analysis*, 12(1):29–45.

Samir Passi and Mihaela Vorvoreanu. 2022. Overreliance on ai literature review. *Microsoft Research*.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. *arXiv preprint arXiv:1906.09821*.

Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. 2019. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2016. Generative deep neural networks for dialogue: A short review. *arXiv preprint arXiv:1611.06216*.

Mehmet Sofi, Matteo Fortier, and Oana Cocarascu. 2022. A robustness evaluation framework for argument mining. In *Proceedings of the 9th Workshop on Argument Mining*, pages 171–180.

Harini Suresh and John V Guttag. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2(8).

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.

Roman Vaculín, Richard Hull, Maja Vukovic, Terry Heath, Nathaniel Mills, and Yutian Sun. 2013. Supporting collaborative decision processes. In *2013 IEEE International Conference on Services Computing*, pages 651–658. IEEE.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jenny S Wesche and Andreas Sonderegger. 2019. When computers take the lead: The automation of leadership. *Computers in human Behavior*, 101:197–209.

H James Wilson and Paul R Daugherty. 2018. Collaborative intelligence: Humans and ai are joining forces. *Harvard Business Review*, 96(4):114–123.

Scott Wright and John Street. 2007. Democracy, deliberation and design: the case of online discussion forums. *New media & society*, 9(5):849–869.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*.

Fabio Massimo Zanzotto. 2019. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, 64:243–252.

Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. 2018. Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8827–8836.

Leveraging High-Precision Corpus Queries for Text Classification via Large Language Models

Nathan Dykes[†] **Stephanie Evert**[†] **Philipp Heinrich**[†]
Merlin Humml[‡] **Lutz Schröder**[‡]
[†]Chair of Computational Corpus Linguistics [‡]Chair of Theoretical Computer Science
Friedrich-Alexander-Universität Erlangen-Nürnberg
[†]Bismarckstr. 6, 91054 Erlangen [‡]Martensstr. 3, 91058 Erlangen
{firstname.lastname}@fau.de

Abstract

We use query results from manually designed corpus queries for fine-tuning an LLM to identify argumentative fragments as a text mining task. The resulting model outperforms both an LLM fine-tuned on a relatively large manually annotated gold standard of tweets as well as a rule-based approach. This proof-of-concept study demonstrates the usefulness of corpus queries to generate training data for complex text categorisation tasks, especially if the targeted category has low prevalence (so that a manually annotated gold standard contains only a small number of positive examples).

Keywords: text categorisation, corpus queries, fine-tuned LLM, argumentation mining

1. Introduction

Gaining an empirical overview of arguments, sentiments, and desires voiced in public discourse is an important prerequisite in technological support for deliberation. Social media have become an increasingly important platform for such publicly voiced opinions, but the automated extraction of computer-mediated natural argumentation is challenging due to the disconnectedness of the statements encountered and the broad variation in their linguistic expression. We work at the boundary of natural language processing, corpus linguistics, argumentation mining, and reasoning in an approach where we use interactively designed corpus queries to capture expressions of relevant phenomena with high precision in a corpus of tweets. In the present contribution, we focus on the possibility of exploiting query matches as training data to fine-tune an LLM, allowing us to increase recall of the queries with only a small loss in precision.

1.1. Related Work

We illustrate our approach to finding argumentative fragments with the running example of expressions of *desire*. The end goal is a formal representation of argumentative statements, leveraging the power of automated reasoners to aid in the difficult task of reconstructing implicit reasoning steps (Boltužić and Šnajder, 2016) and connections between statements (Budzynska and Reed, 2011). The example in this paper belongs to a large inventory of argumentative fragments in our argument mining framework. Each of these fragments represents a concept that we deem relevant to everyday argumentation – besides *desire*, this includes statements about e.g. consequence and group membership. Of course, the presence of *desire* or any other such fragment on its own does not imply the presence of an argument. However, expressions of desire are common building blocks in everyday argumentation and we consider them particularly relevant to deliberation processes.

A straightforward solution for detecting *desire* would be to train a supervised binary classifier on our manually annotated gold standard. Recent work has shown promising results from fine-tuning pre-trained large language models (LLM), which exploits the rich linguistic knowledge encoded in the LLM (see e. g. Rahman et al., 2023; Qiu and Jin, 2024). However, obtaining sufficient training data can still be difficult, especially for complex annotation tasks like our running example: Besides conceptual issues of precisely defining the scope of what is counted as desire, there are many ways to express the concept linguistically. Moreover, the prevalence of *desire* in our data set is low ($\approx$ 6%), so we expect to find only a handful of positive examples even in a relatively large manually annotated gold standard (see Section 2.2). Thus, the automatic identification of such tweets is a challenging task.

In our case study, we compare the approach of fine-tuning an LLM on a manually labelled gold standard to a rule-based approach using manually developed corpus queries developed by (cf. Dykes et al., 2020, 2021). These queries can retrieve thousands of positive examples with high precision, which we can then use as additional training data in fine-tuning the LLM. This combined method

The First Workshop on Language-driven Deliberation Technology (DELITE2024), pages 52–57
20 May, 2024.

outperforms the other approaches by a considerable margin. Our approach thus shares the same goal as *data augmentation*, i. e. "to increase the diversity of training examples without explicitly collecting new data" (Feng et al., 2021, 968). Data augmentation usually adds to a training corpus with artificial examples that are very close to observed instances, or that are developed introspectively. An alternative approach similar to ours is to use "weak labeled data" (Shnarch et al., 2018), where coarse heuristics are applied to extract training examples while allowing for a significant amount of noise. In our approach, we use linguistically sophisticated queries which can extract empirical instances from the overall corpus with high precision to enhance our much smaller manually annotated set.

2. Data and Manual Annotation

2.1. Data

We reconstruct the corpus of Dykes et al. (2020), containing tweets with the token *brexit* (case-insensitive) collected in 2016, i.e. the year of the UK Brexit referendum. We disregard retweets and apply a strict deduplication algorithm (which disregards case shift, @-mentions, URLs, and hashtags). Our data comprises over 4.3 million tweets with approximately 80 million tokens.

Since we also build on the queries from Dykes et al. (2021), we use the IMS Open Corpus Workbench (Evert and Hardie, 2011)[1] for corpus indexing, and apply a similar linguistic annotation pipeline, i.e. Ark TweetNLP (Owoputi et al., 2013)[2] for simple PoS tags, the OSU Twitter NLP tools (Ritter et al., 2011, 2012)[3] for Penn-style PoS tags and named entity recognition, and a lemmatiser based on Minnen et al. (2001). For tokenisation, we use SoMaJo (Proisl and Uhrig, 2016)[4] and reconcile the different tokenisation layers during post-processing.

2.2. Manual Annotation

For manual annotation, two random samples are extracted from the corpus: `pre` consists of 785 of the originally 1000 tweets labelled for *desire* by Dykes et al. (2021) – i.e., the tweets from their study that were still available during our corpus construction. All of these tweets were posted before the Brexit referendum (June 23, 2016). The examples from `pre` are used as a starting point for developing corpus queries (cf. Section 3.1).

[1] https://cwb.sourceforge.io/

[2] http://www.cs.cmu.edu/~ark/TweetNLP/

[3] https://github.com/aritter/twitter_nlp

[4] https://github.com/tsproisl/SoMaJo

		V	E	gold
pre	M	0.627	0.724	0.778
	V		0.579	0.601
	E			0.689
post	M	0.723	0.772	0.906
	V		0.730	0.814
	E			0.890

Table 1: Inter-annotator agreement (kappa scores) for the *desire* pattern.

Since this sample only contains tweets from before the Brexit referendum, we sampled an additional 1000 random tweets posted on August 21, 2016 after the referendum (`post`).[5] Manual annotation of `post` provides additional training data for the LLM and allows us to estimate query *recall* (as unseen test data for the queries).

Additionally, random samples of query matches were annotated to provide reliable estimates of query *precision* (see Section 4). For *desire*, this amounts to a total of 3997 tweets (`matches`). In contrast to `pre` and `post`, this data set is not a random selection of tweets but includes only tweets found by our queries. As it does not show how many instances of *desire* were missed by the queries it cannot be used to reliably estimate *recall*.[6]

Our annotation guidelines are based on those provided by Dykes et al. (2021) and were continuously refined during annotation. For each fragment, we give a description along with positive and negative examples from the corpus. For instance, the description of *desire* differentiates two uses of the word *support*, which is accepted as an expression of desire in *She supports Brexit* but is excluded when referencing actions (*they gave a speech to support Brexit*). Even for human annotators, detecting *desire* is not as straightforward as it may seem intuitively, since it is easily confused with other similar patterns such as the *desirer* pattern (expression of membership in a group of entities desiring a concept, as in *Trump is a Brexit supporter*).

Three student assistants annotated all *desire* statements via a custom web interface. Their annotations were adjudicated regularly, and doubtful cases were discussed with the project members. We report pairwise inter-annotator agreement in Table 1. The kappa scores range from $\kappa = .579$ (direct comparison of annotators V and E on `pre`) to $\kappa = .906$ (agreement of annotator M with the adjudicated gold standard), showing a modest to substan-

[5] Improved deduplication carried out after sampling reduced this data set to 973 tweets.

[6] To put in exaggerated terms: a query with a single true positive and no false positives has a precision of 100%, but this does not say anything about its recall.

tial agreement with the final gold standard. Given the difficulty of the task, we deem these values to be good overall. Our gold standard is available at.[7]

The prevalence of *desire* according to the manual annotation was 4.5% on `pre` and 7.7% on `post`, cf. Table 2; it is thus indeed an infrequent phenomenon.

3. Automatically Detecting *Desire*

3.1. Querying

The queries we use to find further examples of our argumentative fragments are written in the CQP query language (Evert and The CWB Development Team, 2022), enabling complex searches that combine different levels of linguistic annotation.

```
[lemma="all|everything|that|what"]
/entity_np_actor[]
[lemma=$verbs_prefer]
[lemma="be"]
[lemma="for|to"] [pos="DET|A.+"]*
(/entity_np_all[] | [pos="VERB"])
```

The example above is one of 18 queries for *desire* and matches *all ENTITY wants is for/to NP/VP*.[8]

The queries are designed to abstract away from annotated examples as much as possible while maintaining high precision. For instance, because the entity in *desire* statements is almost always a person or an organisation/group, the noun phrase `/entity_np_actor[]` has to contain a proper name or a noun from a manually compiled list of plausible entities.

In total, the queries retrieve 145,699 corpus matches. Table 2 (top) shows the performance of the query approach on our labelled datasets: a recall of 43% on unseen data (`post`), but a very high-precision of 96% (`matches`).

3.2. LLM Fine-Tuning

In this section, we fine-tune an LLM on the binary classification of tweets as to whether they contain *desire*. We consider two models here: firstly, a model trained on a 70% training/test split of the adjudicated gold standard (`combined`, comprising `pre` and `post`). This dataset contains 73 positive and 1158 negative examples. Secondly, a model trained on query matches (excluding matches on `combined` to ensure comparability). We use 70% of all 145,699 matches as positive training examples and add the same amount of random tweets (excluding query matches and those in `combined`). We thus assume all query matches to be instances of *desire* and randomly selected tweets to be negative examples. This is a reasonable approximation since the prevalence of *desire* is ca. 6% and the precision of our queries is ca. 96%.

We opt for distilbert-base-uncased (Sanh et al., 2019) as a base model and fine-tune using the `transformers` package with standard settings. The choice of distilbert-base-uncased for this paper stemmed from its lightweight nature, being nearly half the size of models like bert-base-uncased, its availability off-the-shelf, and the fact that it has shown promising outcomes in prior research (see e. g. Rahman et al., 2023). Although we did explore other models, our experiments consistently demonstrated similar results (see below).

The trained models can be used to calculate scores for both classes (*desire* and *no desire*); we focus on the positive class here. A cut-off value for this score determines the trade-off between precision and recall; Figures 1 and 2 show the resulting precision-recall curves. As a composite measure we use the area under these curves (PR-AUC).

4. Results

Unsurprisingly, the LLM trained on query matches accurately distinguishes query matches from other tweets, despite using 70% of the matches as positive training examples. Evaluation on the remaining 30% (mixed with random tweets) yields a PR-AUC of 0.9978. However, we are interested in its performance to detect *desire* in general, not limited to instances that are also found by the queries (whose estimated recall is only 43%).

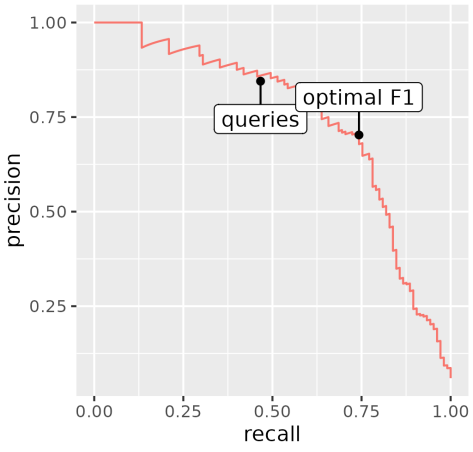


Figure 1: PR curve of LLM trained on query matches and evaluated on `combined`.

The PR curve of this LLM on `combined` (Figure 1) shows that decent trade-offs between precision and recall are possible. It is no coincidence

[7] Link will be provided with the final submission.

[8] The query was slightly shortened for the sake of readability. Due to limited space, we cannot provide a detailed explanation of the query syntax.

data set	prev.	approach	FN	FP	TN	TP	precision	recall	F_1
`pre`	0.08		31	3	721	30	0.91	0.49	0.64
`post`	0.05	queries	25	6	923	19	0.76	*0.43*	0.55
`matches`	0.96			94		2312	*0.96*		
`combined`	0.06	LLM (matches)	28	33	1620	77	0.70	**0.73**	**0.72**
	0.06	queries	56	9	1644	49	**0.84**	0.47	0.60
test-split	0.06	LLM (matches)	9	6	489	23	0.79	**0.72**	**0.75**
	0.06	LLM (combined)	19	26	469	13	0.33	0.41	0.37
	0.06	queries	17	2	493	15	**0.88**	0.47	0.61

Table 2: Top: Evaluation of corpus queries for *desire* on different data sets. Recall can most reliably be estimated from `post`, while precision can most reliably be estimated on actual query `matches` (indicated in italics). Middle and bottom: comparison of different approaches on the complete data set `combined` (middle) and on the test split of `combined` (bottom). The query approach yields the highest precision, and the LLM trained on query matches yields the highest recall (indicated in bold).

that the performance of the queries themselves lies on this curve: The LLM can near-perfectly retrieve query results and at this point, its predictions are almost identical to the query matches. Moving down the PR curve, we buy recall by spending precision. We also indicate the optimal cut-off point maximising F_1, i.e. the harmonic mean between precision and recall. We determine this value *ex post* for reasons of simplicity, but it could also be determined on a separate development set.

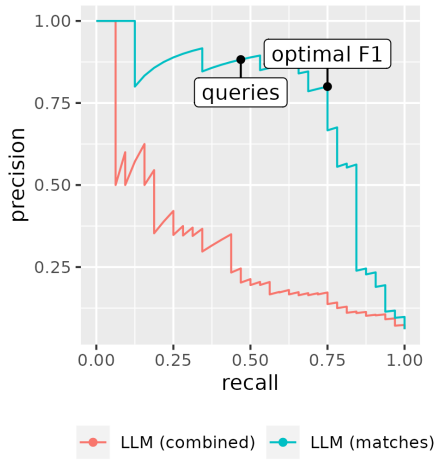


Figure 2: PR curves on test-split of `combined`.

Figure 2 evaluates both trained LLMs on the test split of `combined`. The LLM trained only on 73 positive and 1158 negative examples performs poorly in comparison to the LLM trained on query matches. Table 2 lists detailed results for all approaches on `combined` and its test split (for LLMs, the numbers shown are taken at the point of optimal F_1). In terms of precision, the queries yield the best results (as by design). However, the LLM trained on query matches can yield better recall, as is exemplified by the point of optimal F_1 on the PR-curve.

5. Discussion

Examining tweets that are true positives (TP) of the LLM at the point of optimal F_1 but not found by the corpus queries shows that the higher recall of the LLM approach can be attributed to several interpretable factors:

Most new TPs contain typos (*Britian*) or short insertions (*Denmark for one will be queuing up to leave*). While the queries could likely be adjusted to find such cases, this would either introduce unnecessary complexity or compromise precision.

Other new TPs are due to errors in the linguistic pre-processing used by queries, e.g. several nominalised adjectives that were incorrectly treated as adjectives by the PoS tagger and thus not found by queries (*The British want EU migrants to stay*). Similarly, the queries impose semantic restrictions via wordlists. The LLM, on the other hand, also finds instances of *desire* with unusual entities such as *noted Europhile paper backs Brexit*.

Finally, the LLM found some tweets with syntactic patterns for which no queries had been written – either because the expression contained non-standard syntax (*If we Brexit., ending the Barnet agreement, I'm for!*), or because the constructions were too rare to reasonably justify developing a manual query (*Very much looking forward to seeing nigel farage in action tonight*).

Most false positives (FP) of the LLM, which were not matched by the queries, are syntactically similar to one of the queries without expressing the correct semantics (*#Brexit gloom is for losers*). Fewer tweets allude to desire more implicitly than allowed by the guidelines (*"Being pro brexit is wacist!" said the hipster white brits to the black brits* – this tweet is not accepted because it is a general statement rather than a specific entity desiring something).

6. Conclusion

In conclusion, manually engineered corpus queries can retrieve argumentative fragments with very high precision but limited scalability. Tweets containing typos or unusual constructions are often missed. Using an LLM fine-tuned on query results, on the other hand, allows us to choose the trade-off between precision and recall freely along the PR curve. Compared to the query matches, the LLM can retrieve considerably more relevant tweets. Based on the new TPs found in the gold standard, the additional hits can also be expected to reflect some of the typical CMC features that are often filtered out by the queries.

Note that considerable improvements of the LLM predictions are quite possible. Firstly, training on all query results could be explored, but would no longer allow us to assess the LLM's ability to predict query results. Secondly, using a data set with the estimated prevalence of *desire* for training could be beneficial. Lastly, experimenting with different base models and hyperparameter settings (such as learning rate, weight decay, etc.) is another avenue. However, our primary objective here was to establish a proof of concept rather than engineering an optimal system.

Acknowledgements

The work is supported by the DFG under Project 377333057 *Reconstructing Arguments from Newsworthy Debates (RAND)*.

7. Bibliographical References

Filip Boltužić and Jan Šnajder. 2016. Fill the gap! analyzing implicit premises between claims from online debates. In *Argumentation Mining*, 3.

Katarzyna Budzynska and Chris Reed. 2011. Speech acts of argumentation: Inference anchors and peripheral cues in dialogue. In *Proceedings of 11th International Conference on Computational Models of Natural Argument (CMNA 2011)*.

Natalie Dykes, Stefan Evert, Merlin Göttlinger, Philipp Heinrich, and Lutz Schröder. 2020. Reconstructing Arguments from Noisy Text: Introduction to the RANT project. *Datenbank-Spektrum*, 20:123–129.

Natalie Dykes, Stefan Evert, Merlin Göttlinger, Philipp Heinrich, and Lutz Schröder. 2021. Argument parsing via corpus queries. *it - Information Technology*, 63(1):31–44.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Corpus Linguistics, CL 2011*. University of Birmingham.

Stefan Evert and The CWB Development Team. 2022. *The IMS Open Corpus Workbench (CWB) CQP Interface and Query Language Tutorial*. CWB Version 3.5.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Nat. Lang. Eng.*, 7(3):207–223.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Human Language Technologies, HLT-NAACL 2013*, pages 380–390. ACL.

Thomas Proisl and Peter Uhrig. 2016. SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 57–62, Berlin. Association for Computational Linguistics.

Yunjian Qiu and Yan Jin. 2024. Chatgpt and fine-tuned bert: A comparative study for developing intelligent design support systems. *Intelligent Systems with Applications*, 21:200308.

A M Muntasir Rahman, Wenpeng Yin, and Guiling Wang. 2023. Data augmentation for text classification with EASE. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 324–332, Online. Association for Computational Linguistics.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Empirical Methods in Natural Language Processing, EMNLP 2011*, pages 1524–1534. ACL.

Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from twitter. In *Knowledge Discovery and Data Mining, KDD 2012*, pages 1104–1112. ACM.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled

version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

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

8. Language Resource References

All language resources used in our research (NLP tools and LLM) have accompanying references. We prefer to cite them in this way rather than as language resources in order to give authors proper credit in citation metrics.

Author Index