Predicting Moderation of Deliberative Arguments: Is Argument Quality the Key?

Neele Falk*, Iman Jundi*, Eva Maria Vecchi, Gabriella Lapesa
Institute for Natural Language Processing
University of Stuttgart (Germany)
first[-middle].last@ims.uni-stuttgart.de

Abstract

Human moderation is commonly employed in deliberative contexts (argumentation and discussion targeting a shared decision on an issue relevant to a group, e.g., citizens arguing on how to employ a shared budget). As the scale of discussion enlarges in online settings, the overall discussion quality risks to drop and moderation becomes more important to assist participants in having a cooperative and productive interaction. The scale also makes it more important to employ NLP methods for (semi-)automatic moderation, e.g. to prioritize when moderation is most needed.

In this work, we make the first steps towards (semi-)automatic moderation by using state-of-the-art classification models to predict which posts require moderation, showing that while the task is undoubtedly difficult, performance is significantly above baseline. We further investigate whether argument quality is a key indicator of the need for moderation, showing that surprisingly, high quality arguments also trigger moderation. We make our code and data publicly available.[1]

1 Introduction

Digital innovations reshaped direct democracy, allowing a large group of citizens to be involved in public decisions. However, as pointed out by Lampe et al. (2014), "Participation in discussions about the public interest can be enhanced by technology, but can also create an environment in which participants are overwhelmed by the quantity, quality, and diversity of information and arguments.". Poor quality discussions risk to lead to poor decisions. That is why research on Deliberative Theory does not focus on the output of democratic decision-making, but on the discourse exchange that precedes it (Bächtiger and Parkinson, 2019; Steenbergen et al., 2003; Steiner et al., 2005).

Moderation plays a crucial role in deliberation, in presence or online. Moderators assist participants in having a cooperative and productive discussion: indeed, in the deliberation terminology, moderation is referred to as *facilitation* (Kaner et al., 2007; Trénel, 2009), highlighting the positive nature of the moderation outcome more than the negative assessment of its input. A closer look at moderation in deliberative settings reveals a quite heterogeneous set of actions: stimulate discussion, solve conflicts, help participants to formulate clear and understandable comments while keeping them on topic and focus on solutions, make participants feel valued (e.g., addressing them personally).

The need for trained moderators is a clear bottleneck when scaling to large audiences. NLP can provide a crucial contribution, by making moderation automatic or (more realistically) semi-automatic – providing support to human moderators in some of the many moderator tasks. An important task here could be defined as identifying when a moderation intervention is needed. Our work makes the first steps towards NLP moderation and this task, by modeling and analysing the moderation signal in an e-deliberation forum *RegulationRoom* (Park and Cardie, 2018), addressing two research questions: (Q1) Can we predict automatically whether a forum post (comment) will trigger a moderation intervention? (Q2) What is the relation between Argument Quality (Wachsmuth et al., 2017a,b) and moderation and is AQ a key signal for moderation?

Regarding (Q1) we show that state of the art classification models can predict moderation reasonably well, but the task is quite difficult and even taking into account a broader context (i.e., preceding post) does not significantly improve performance. The moderation signal might be too heterogeneous, suggesting that the different moderation functions

[*] denotes equal contribution

[1] Code and annotated sample available here: https://github.com/imanjundi/arguments-moderation

Proceedings of The 8th Workshop on Argument Mining, pages 133–141
Punta Cana, Dominican Republic, November 10–11, 2021. ©2021 Association for Computational Linguistics

might be better targeted individually. We thus focus on a subset of the *RegulationRoom* dataset annotated with moderator functions, and concentrate on those explicitly annotated as an attempt to improve comment quality. To address (Q2) we employ an off-the-shelf Argument Quality classifier (Lauscher et al., 2020) and conduct a manual annotation experiment to analyse the relation between moderation and AQ. This can help us better understand if low-quality comments are mainly the ones that trigger moderation; the results are interesting and surprising: contrary to intuitive expectations, high AQ also triggers moderation. In high AQ cases, it seems that moderators ”pick up” on interesting, well made points and ask the users to elaborate further.

The contribution of this work is thus twofold: first, we address the task of moderation prediction for the first time and with encouraging results; second, we uncover unexpected dynamics of the modulation of AQ with respect to moderation, thereby contributing to the empirical characterization of the moderation signal in deliberative settings.

2 Data

We investigate our research questions on data taken from the deliberation platform, *RegulationRoom*[2] which allows ordinary citizens to take part in the decision process about regulations proposed by federal agencies. The agencies provide detailed information about the rules under discussion such that participants can comment and share their opinions about them. To promote a fruitful discussion, human moderators monitor the discussion and intervene to help and support the participants. Besides general supervision functions (policing, helping with technical difficulties) the moderators can intervene to help users improving the content of their posts or to activate other participants to join the discussion. The moderators were trained and equipped with a ‘moderator protocol’ (eRulemaking Initiative et al., 2017), which describes the possible reasons for an intervention. Figure 1 illustrates an example of a moderator intervention in *RegulationRoom*.

The full *RegulationRoom* dataset (used in the experiments in section 3.1) contains 3k comments spanning various topics (refer to Table 3 in the Appendix for the distribution of the different topics) with 717 (23.63%) moderated comments vs. 2317 (76.36%) not-moderated.

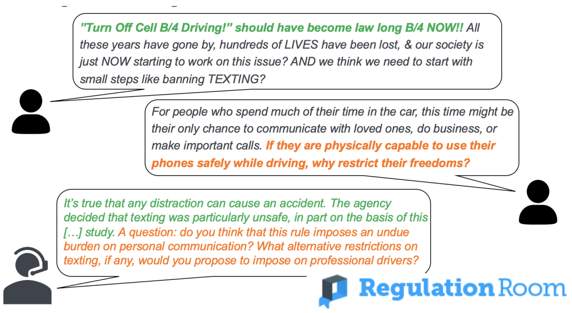

”Turn Off Cell B/4 Driving!” should have become law long B/4 NOW!! All these years have gone by, hundreds of LIVES have been lost, & our society is just NOW starting to work on this issue? AND we think we need to start with small steps like banning TEXTING?
For people who spend much of their time in the car, this time might be their only chance to communicate with loved ones, do business, or make important calls. If they are physically capable to use their phones safely while driving, why restrict their freedoms?
It’s true that any distraction can cause an accident. The agency decided that texting was particularly unsafe, in part on the basis of this […] study. A question: do you think that this rule imposes an undue burden on personal communication? What alternative restrictions on texting, if any, would you propose to impose on professional drivers?
Regulation Room

Figure 1: Moderation in RegulationRoom

A subset of moderator interventions from this data (303 comments, topics marked with an asterisk in Table 3) has been annotated in Park et al. (2012) with the reasons for an intervention (moderator functions). The annotation schema for the moderator functions was inspired by the original moderator training protocol but due to the large variety of different possible reasons for a moderator intervention, the functions where merged into more general categories. The authors use the annotated subset to predict what type of moderator action would be required, focusing on the two most frequent types of intervention, *broadening the discussion* and *improving argument quality* (refer to Table 5 in the Appendix for a full list of annotated categories and their distribution). Unlike Park et al. (2012), we focus on predicting and analyzing when a moderation intervention is needed which was still not tackled before, although this is where the bottleneck in online moderation usually lies.

3 Experiments

We conduct classification experiments on two tasks: predicting if there was a moderator intervention for a comment *Any Moderation Intervention* then narrowing the scope to predicting if there was a moderator intervention to improve quality for a comment *Quality Moderation Intervention*. We fine-tune a Pretrained Language Model, RoBERTa-base (Liu et al., 2019), with a classification head using the Huggingface Transformers library (Wolf et al., 2020) for each task.

The dataset for each experiment is split into 60%, 20%, 20% train, dev and test respectively. Given the small size of the dataset, we evaluate on a 5-fold split of the whole dataset and report the average and standard deviation to get a more robust result that is less sensitive to the variance caused by the random dataset split. We compare the trained mod-

[2] http://regulationroom.org/

context	BAC	F1 macro	F1 positive
(random)	51.08 ± 2.2	47.70 ± 2.5	33.37 ± 3.0
parent=0	64.77 ± 2.1∗	62.51 ± 2.2∗	46.32 ± 2.6∗
parent=1	65.30 ± 1.1	63.63 ± 1.4	47.07 ± 1.4
parent=2	66.03 ± 1.7	64.95 ± 1.6	47.99 ± 2.5

Table 1: Results of *Any Moderation Intervention*
The difference is statistically significant ($p < 0.05$)

context	BAC	F1 macro	F1 positive
(random)	50.14 ± 1.3	45.40 ± 2.1	28.99 ± 1.5
parent=0	58.52 ± 3.5∗	56.98 ± 2.6∗	34.27 ± 5.0∗
parent=1	58.14 ± 6.3	56.64 ± 6.9	34.92 ± 8.1
parent=2	55.38 ± 4.3	55.41 ± 4.1	27.44 ± 7.1

Table 2: Results of *Quality Moderation Intervention*
The difference is statistically significant ($p < 0.05$)

els to a random prediction baseline and report the balanced accuracy (BAC), F1 macro and F1 for the positive class. The statistical significance is calculated using McNemar's Test.

To represent the context of the comment we consider the previous $parent \in \{0, 1, 2\}$ comments where $parent = 0$ is considering the comment itself alone with no previous context. We concatenate the comments to form the full context while adding a separation token </s> in-between.

We train for 10 epochs and choose the best performing model on the dev split, finally reporting the results on the test split. We do hyperparameter optimization on the dev split using grid search over $parent \in \{0, 1, 2\}$ and learning rate $lr \in [3e^{-6}, 3e^{-5}]$. We also try sequence length 256 and 512 to deal with the increased length due to the longer context with more parent comments. We further do a search for the optimal threshold (threshold moving) to better deal with data imbalance.

3.1 Any Moderation Intervention

In our first experiment we train a model to predict any moderator intervention. This has practical application in automatically assisting moderators and predicting where their effort is most needed. This is particularly crucial with moderation on a large-scale where it's not feasible for the moderator to check every comment.

The full dataset (Table 3) is used here where the moderator intervention is used as a noisy label: a user comment followed by a moderator comment is labeled as moderated, 23.63% of the comments, while the rest is labeled as not-moderated. The model performs better than chance as seen in in Table 1, but it is also clear that the task is still difficult. This can be due to the noisy labels, the ambiguity inherent in the task, and the small size of the training data.

Including more context leads to better performance although the increase is not statistically significant, but that might be also due to the small size of the dataset. Consider, however, that the majority of the comments (∼60%) is not nested i.e. it is only for ∼40% that we can exploit a larger context to support classification prediction (see Table 4 in the Appendix for more details).

3.2 Quality Moderation Intervention

In our second experiment we narrow the task to predicting a moderator intervention for improving a comments quality. This type of moderation is one of the most important functions in a deliberative setting, as good quality contributions help to ensure that the different perspectives are understood by others. While it is questionable whether a function like *broadening discussion* can be predicted from the content of the discussion post itself, it is expected that its quality can at least be approximated from certain linguistic and structural particularities in the comment itself. If this is the case, this type of moderation is also suitable to be supported by automated methods. By dropping comments that were moderated for other reasons, the positive instances may form a more homogeneous group and may therefore be easier to model.

For this experiment we rely on the annotated subset (Park et al., 2012) described in section 2, considering only comments whose moderator intervention was annotated with *improve comment quality* as positive instances. We use all other comments from the same topics as negative instances. This subset (further referred to as *RegroomForQual*) contains 876 negative and 222 positive instances (∼ 75%/ ∼ 25%). Again, the model performs better than the baseline (cf. Table 2), but has a lower performance than in the first experiment. Adding context does not improve the results but rather hurts the performance (the distribution of available context in *RegroomForQual*, displayed in Table 6 in the Appendix is comparable to that of the larger dataset).

4 Analysis

Given that the task of predicting *Quality Moderation Intervention* is still challenging to model, this raises the question of how exactly *comment quality* was defined by the moderators/annotators of the underlying dataset and to what extent an explicit prediction of, for example argument quality, would help to improve the results on this task. To gain a more comprehensive understanding of the argument quality in this dataset, we collected quality scores using a manual and an automatic annotation to examine them in the context of a moderation intervention.

4.1 AQ Annotation

In both research areas, argument mining and deliberative theory, theoretical and empirical definitions of argument quality exist. They range from assessing more specific aspects of argument quality (e.g. persuasion in Liane Longpre and Cardie, 2019) to measuring argument quality as a relative assessment (given two arguments, which is more convincing (Habernal and Gurevych, 2016). In deliberative theory the focus lies on aspects important for a fruitful discussion related to a decision process (e.g. respectful tone, constructiveness) and aggregating the single scores to measure the quality of the overall discourse.

Manual Annotation Although there is a large variance in different definitions of quality and methods to assess it, it was shown that a common perception of overall argument quality exists and that there are correlations between different measurement methods (Wachsmuth et al., 2017a). We therefor simplify our annotation for this phenomenon and quantify argument quality in one aggregated score. This allows us to check whether the annotators' intuitive understanding of Argument Quality is consistent with the moderator interventions. We still provide the annotators with detailed information on sub-dimensions of argument and deliberative quality to consider when annotating[3].

We collected a set of AQ annotations on comments sampled from the *RegroomForQual* dataset. Four annotators were asked to estimate the argument quality of 112 comments, providing a score ranging from 0 (very low quality) to 5 (very high quality). Note that the annotators did not know whether a comment triggered a moderator intervention or not.

The items were sampled balancing two criteria: moderated vs. not moderated; high quality vs. low quality according to the off-the-shelf AQ model by Lauscher et al. (2020).[4] The average pairwise inter-annotator agreement (weighted Cohen's kappa) was 0.40, and the average pairwise correlation between annotators was 0.62 (pairwise details in Table 7).

Automatic Annotation We use a classifier by Lauscher et al. (2020) to automatically score the comments on the three core argument quality dimensions *Cogency, Effectiveness, Reasonableness* (Wachsmuth et al., 2017b) and an aggregated score (*overall quality*). The classifier was trained on the *GAQCorpus* (Ng et al., 2020) that contains data from different online forums annotated with scores between 1 and 5. As a sanity check, we first compared the human AQ on our annotated set to the automatically predicted ones: AQ scores predicted by the system correlate with human ones ($\tau = 0.72, p < 0.01$ with *overall quality*, $\tau = 0.70, p < 0.01$ with all three subdimensions). The scores for the subdimensions highly correlate with *overall quality* ($\tau > 0.98, p < 0.01$), so we focus on this score for further investigations.

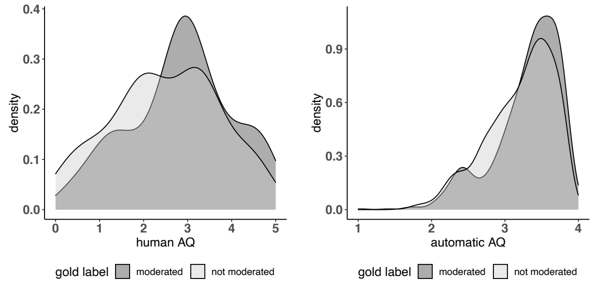
4.2 Relation between Moderation and AQ

To get a first impression about the relationship between moderation and argument quality we look at the distributions of the quality scores for the comments that triggered moderation vs. the comments that did not trigger an intervention. The intuitive expectation is that comments with low quality scores would be more likely to trigger moderation to improve their quality. The visualization of the actual distributions in Figure 2 does not confirm this hypothesis. For the manually annotated data, we find a peak in the middle range of scores (2 to 4), but both lower and higher quality comments were moderated (see Fig. 2a). Likewise, there were also low quality comments that were not moderated. The distributions for the automatic scores initially show that they have less variance (few comments were annotated with extreme scores), with a tendency to mid-high scores (most data points have scores between 3 and 4, maximum score would be 5). There is no significant difference between the distributions of moderated and non-moderated data, which again disproves the hypothesis that low-quality comments are more likely to be moderated (cf. Fig. 2b).

Similarly we analyse the model that was trained

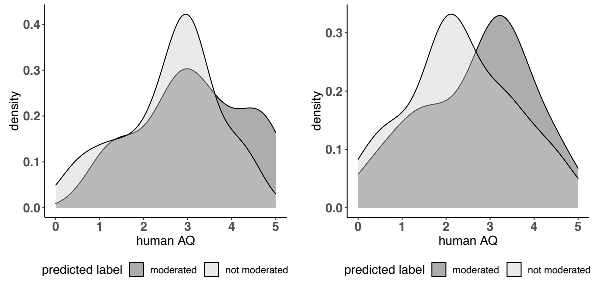
[3] The Annotation Guidelines are provided in the Appendix.

[4] We provide all details on the sampling criteria for HQ vs. LQ in the Appendix, section A.4.1.



(a) Annotation with human AQ (b) *RegroomForQual* with automatic AQ

Figure 2: Density plots for AQ: Comparing distribution for moderated and not moderated comments



(a) True positives vs. false negatives (b) True negatives vs. false positives.

Figure 3: Density plots for AQ: Comparing distribution for model predictions, using the best model trained on *Quality Moderation Intervention*.

on *quality moderation intervention* to check if it picks up on Argument Quality as a signal for an intervention. We plot the distribution of the human annotated quality scores for the moderated instances (gold label = 1) and compare the true positives and false negatives (Figure 3a): The model is able to capture the majority of the high-quality comments and struggles to correctly classify comments with medium quality scores (2-3). For the instances that were not moderated (gold label = 0, Figure 3b), the model produces a high amount of false positives for arguments with high AQ, so we can see that for both types of comments, moderated and not moderated, the model is more likely to predict a moderator intervention for higher quality arguments (cf. Figure 3).

We notice a similar pattern on the *RegroomForQual* data: moderation is triggered by both low and high AQ. The probabilities of the model for a moderator intervention correlate positively with the automatic quality scores ($\tau = 0.33, p < 0.01$).

This trend shows that a simplified definition of AQ is not the key to moderation. Neither the automatically produced AQ scores, nor the annotators' intuition of high and low AQ do correlate with moderators' interventions. While agreement is most often encouraging, the disagreement among annotators (e.g. annot3 and annot4 in Table 7) suggests that AQ may be a hard notion to pinpoint.

In fact, there are various cases which are quite hard to annotate based on a general notion of argument quality. In some cases, an argument is well-written, clearly presented, and properly supported with examples or data, however the argument itself is not good and should be considered for moderation. Examples include arguments based on personal accounts (hearsay/antidote) and conspiracy theories (false or intentionally contrived support), see Appendix A.5 for specific examples. In these cases AQ is not only hard to assess, even if measured based on a more fine-grained definition of AQ (e.g. several scores on different aspects of quality), but also these types of arguments might score high on the majority of dimensions and still be moderated in a deliberative setting. Furthermore moderating high-quality arguments can be beneficial if they present a new opinion to help other participants understand that perspective. Moderation is thus a multidimensional, complex problem for which a sole focus on argument quality seems to not be sufficient.

5 Conclusion

Our experiments target the prediction of the need for a moderator intervention in a deliberative setup: our models perform better than chance, but the task is very challenging due to the ambiguity inherent in the task and to the small data available usually in such domains. We further explore the interplay of moderation and AQ (automatically predicted and manually annotated): contrary to the expectation of moderation exclusively ”correcting” low quality, high quality is very often moderated, as well.

Acknowledgments

We would like to thank Anne Lauscher for providing us with the code and models for the AQ classifier. We acknowledge funding by the Bundesministeriumfur Bildung und Forschung (BMBF) through the project E-DELIB (*Powering up E-deliberation: towards AI-supported moderation*).

References

Andre Bächtiger and John Parkinson. 2019. *Towards a New Deliberative Quality*. Oxford University Press, Cambridge, MA, USA.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Cornell eRulemaking Initiative et al. 2017. Ceri (cornell e-rulemaking) moderator protocol.

Sam Kaner, Lenny Lind, Catherine Tolid, Sarah Fisk, and Duane Berger. 2007. *Facilitator's guide to participatory decision-making*. John Wiley & Sons/Jossey-Bass, San Francisco.

Cliff Lampe, Paul Zube, Jusil Lee, Chul Hyun Park, and Erik Johnston. 2014. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly*, 31(2):317 – 326.

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4563–4574, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Esin Durmus Liane Longpre and Claire Cardie. 2019. Persuasion of the undecided: Language vs. the listener. In *Proceedings of the 6th workshop on Argument Mining*, pages 167–176.

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

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

Joonsuk Park and Claire Cardie. 2018. A corpus of eRulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Joonsuk Park, Sally Klingel, Claire Cardie, Mary J. Newhart, Cynthia Farina, and J. Vallbé. 2012. Facilitative moderation for online participation in erulemaking. In *dg.o '12*.

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

Jürg Steiner, André Bächtiger, Markus Spörndli, and Marco R. Steenbergen. 2005. *Deliberative Politics in Action: Analyzing Parliamentary Discourse*. Theories of Institutional Design. Cambridge University Press.

Matthias Trénel. 2009. Facilitation and inclusive deliberation. *Online deliberation: Design, research, and practice*, pages 253–257.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017a. Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255, Vancouver, Canada. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017b. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Appendix

A.1 *RegulationRoom* full dataset: statistics

Topic	count	percentage
Consumer Debt Collection	971	32.00%
Airline Passenger Rights (*)	931	30.68%
Home Mortgage	235	7.74%
NYC Congestion	207	6.82%
Health IT	174	5.73%
Electronic On-Board Recorders (*)	167	5.50%
Proposed Move NY Fair Plan	163	5.37%
Air Travel Accessibility	128	4.21%
Distracted Driving & Texting	30	0.98%
Social Media in Rulemaking	28	0.92%
Total	3034	

Table 3: Moderation dataset: statistics. Full dataset is used in *Any Moderation Intervention* experiments. The (annotated) subtopics marked with * are used in the *Quality Moderation Intervention* experiments

context	count	percentage
parent=0	1850	60.96%
parent=1	604	19.90%
parent=2	485	15.98%
parent>2	95	3.16%

Table 4: *RegulationRoom* Moderation Dataset: Available Context

A.2 *RegulationRoom* subset by (Park et al., 2012): moderator functions

moderator intervention type	frequency
social functions	227
resolving site use issues	18
organizin discussion	19
policing	1
keeping discussion on target	19
improving comment quality	238
broadening discussion	84

Table 5: Moderator intervention types in the *RegulationRoom* subset by (Park et al., 2012)

A.3 *RegRoomForQual* subset of (Park et al., 2012)

context	count	percentage
parent=0	620	56.47%
parent=1	211	19.21%
parent=2	177	16.12%
parent>2	90	8.20%

Table 6: *RegRoomForQual* subset: Available Context

A.4 Annotation study: details

A.4.1 Sampling criteria for HQ vs. LQ

To sample the HQ vs. LQ items for our annotation, we proceeded as follows:

1. We annotated the comments in *RegRoomForQual* with the AQ classifier by Lauscher et al. (2020). This model is trained on a mixed-domain corpus (Ng et al., 2020) including online forum posts and is based on a more coarse-grained version of the taxonomy for AQ by (Wachsmuth et al., 2017a).

2. The distribution of the automatic AQ annotations is as follows: minimum value: 1.7; 1st quartile: 2.9; median: 3.3; mean: 3.2; 3rd quartile: 3.6; maximum value: 3.9. Note that while the range of the classifier is potentially between 1 and 5, we only cover a subportion of that range.

3. We exclude the 10 percent of the shortest and the 10 percent of the longest comments from the sampling. The comments have a length between 29 and 319 tokens, with a mean of 129.

4. We sampled 56 posts (25 moderated, 25 not moderated) from the highest quartile (HQ) and 56 (25 moderated, 25 not moderated) from the lowest quartile (LQ)

A.4.2 Annotators and Agreement

Out of the four annotators, three are MA students in Computational Linguistics who had completed an Argument Mining course. One of them is a native speaker of English, the other two have an excellent command of it. The fourth annotator is one of the authors of the paper, and a native speaker of English.

Table 7 displays the pairwise weighted κ and Pearson correlation for the four annotators, along with their average values.

	κ_w	r
annot 1 / annot 2	0.55	0.76
annot 1 / annot 3	0.26	0.52
annot 1 / annot 4	0.56	0.75
annot 2 / annot 3	0.32	0.53
annot 2 / annot 4	0.46	0.68
annot 3 / annot 4	0.23	0.47
average	0.40	0.62

Table 7: Agreement btw. annotator pairs and average: weighted Cohens κ and Pearson correlation.

A.5 Annotation Examples: difficult cases

Below are a set of examples of comments that follow either a “conspiracy theory” approach or are personal accounts, which in the end are difficult to annotate based entirely on argument quality.

- Example 1 This is merely another phase of Big Brother. For all we know, they could have listening devices installed within. One thing about it: If the door is opened even merely enough for someone to get their foot inside, they always tend to take more and more. That is exactly how this government has grown so out of control. Now they are financing their purses with more money which could have been better spent in maintenance of our vehicles or the payment of taxes we already owe. If you just think about it: why did they just go through hiring so many new law enforcement officers? Law Enforcement needs to focus more on those who PURSUE making money illegally, and let a person earn a fair income honorably.

  human annot.: 2 ; 0 ; 2 ; 2 ; (mean = 1.5)
  moderated: no
  automatic AQ: 3.5

- Example 2 Airlines should require a gate agent to give hourly status or tell passengers when to return to the gate for status. I endured an all-day wait without being advised of status. We were scheduled for an 8am departure. Gate personnel advised us of mechanical problems and said to check back at 10am. At 9am, a gate agent advised maybe a 11am boarding (people who did not arrive till 10am per prior instruction never heard that announcement). After that NO ONE gave us status. It was 6pm before we found out the flight was canceled by calling the airlines. The notice on the departure boards and at the gate said “delayed”... a gate agent never gave us status after 9am! A fellow passenger happened to work at this airline, so he made some calls and was able to get status from the maintenance crew... the maintenance chief said his estimated time of fix was 5PM - maybe. If some “official” from the airline had done the same thing, the passengers could have done something else for those many hours rather than hang around the airport, checking the departures board for status. Also no compensation given...but that’s another discussion. No, it would not have changed the delay/cancellation situation, but if an agent had said check back in x number of hours, then actually come to the gate at that time, people could have done something else. I’ve worked in customer service before. Customers want to know what’s happening if only to be told “this is the problem, we’re still working on it...”

  human annot.: 3; 3; 5; 2; (mean = 3.25)
  moderated: no
  automatic AQ: 3.7

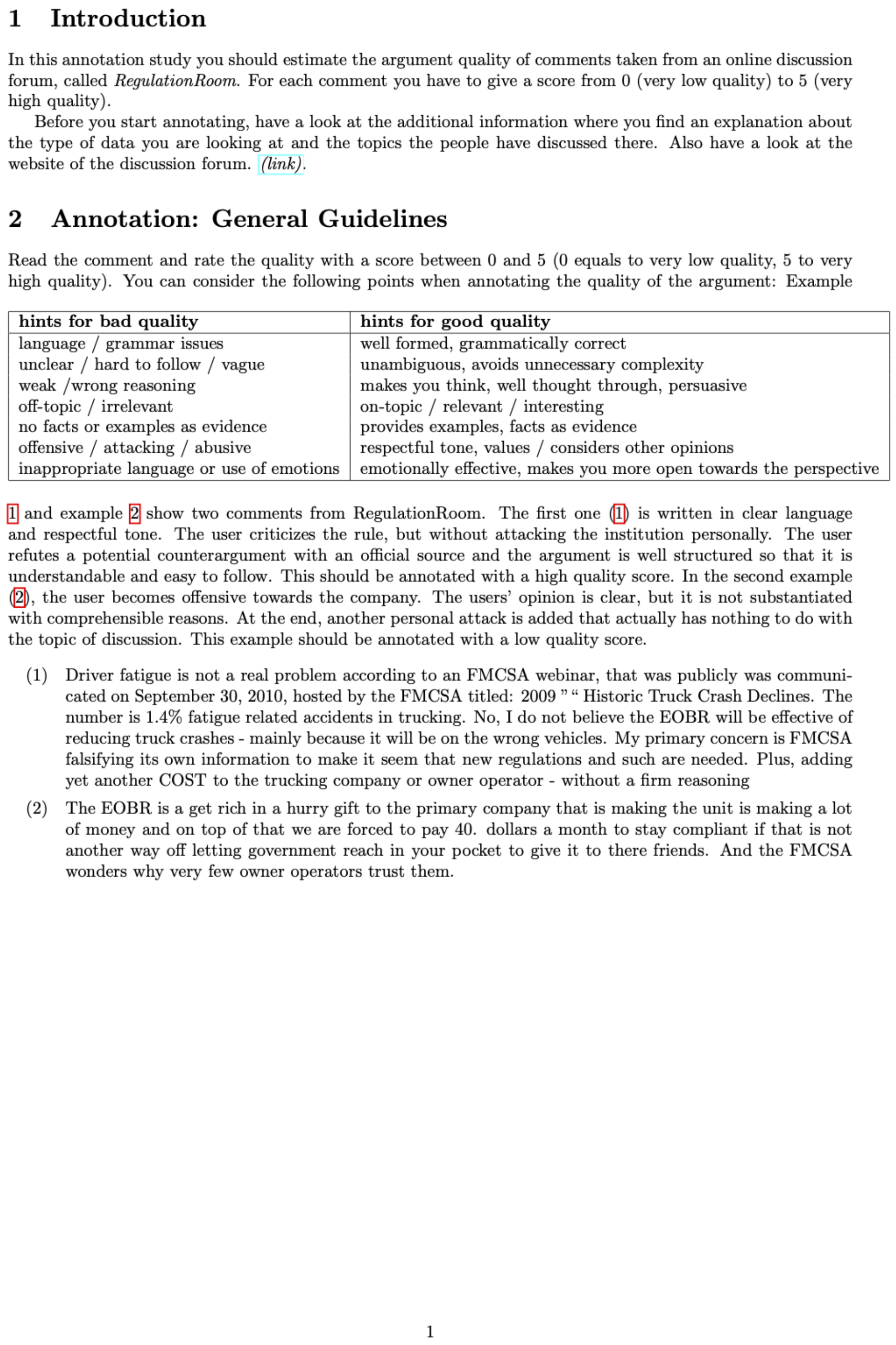
- Example 3 I do not think there should be a compensation cap. If a passenger is bumped from a flight and misses an important business meeting, high school graduation or wedding, there is no price that can compensate for that. Certainly not $1300. In addition, even if the entire flight fare is refunded (easily over $1300 on some long overseas flights), there are other costs that the bumped passenger could face such as non-refundable/prepaid hotel reservations or other travel and travel related expenses. If a passenger can prove that he or she has additional expense above and beyond what the airline offers and the government requires he or she should be awarded more compensation.

  human annot.: 5; 5; 5; 4; (mean = 4.75)
  moderated: yes
  automatic AQ: 3.7

  Moderator response: Note that the proposed regulation states that for delays longer then 2hrs, a passenger can receive compensation amounting to double the price of their ticket or up to $1300. I understand that you think that the cap is too low, but does this doubling of the price of the ticket address your concern that non-refundable expenses should be paid for as well? Or would you propose something like double the ticket price and expenses on top of that? At what point, if any, do you think it becomes unfair for the airline to have to pay these expenses?

A.6 Annotation guidelines

The Annotation Guidelines are depicted in Figure 4.



1 Introduction

In this annotation study you should estimate the argument quality of comments taken from an online discussion forum, called *RegulationRoom*. For each comment you have to give a score from 0 (very low quality) to 5 (very high quality).

Before you start annotating, have a look at the additional information where you find an explanation about the type of data you are looking at and the topics the people have discussed there. Also have a look at the website of the discussion forum. *(link)*.

2 Annotation: General Guidelines

Read the comment and rate the quality with a score between 0 and 5 (0 equals to very low quality, 5 to very high quality). You can consider the following points when annotating the quality of the argument: Example

hints for bad quality	**hints for good quality**
language / grammar issues	well formed, grammatically correct
unclear / hard to follow / vague	unambiguous, avoids unnecessary complexity
weak /wrong reasoning	makes you think, well thought through, persuasive
off-topic / irrelevant	on-topic / relevant / interesting
no facts or examples as evidence	provides examples, facts as evidence
offensive / attacking / abusive	respectful tone, values / considers other opinions
inappropriate language or use of emotions	emotionally effective, makes you more open towards the perspective

1 and example 2 show two comments from RegulationRoom. The first one (1) is written in clear language and respectful tone. The user criticizes the rule, but without attacking the institution personally. The user refutes a potential counterargument with an official source and the argument is well structured so that it is understandable and easy to follow. This should be annotated with a high quality score. In the second example (2), the user becomes offensive towards the company. The users' opinion is clear, but it is not substantiated with comprehensible reasons. At the end, another personal attack is added that actually has nothing to do with the topic of discussion. This example should be annotated with a low quality score.

(1) Driver fatigue is not a real problem according to an FMCSA webinar, that was publicly was communicated on September 30, 2010, hosted by the FMCSA titled: 2009 " " Historic Truck Crash Declines. The number is 1.4% fatigue related accidents in trucking. No, I do not believe the EOBR will be effective of reducing truck crashes - mainly because it will be on the wrong vehicles. My primary concern is FMCSA falsifying its own information to make it seem that new regulations and such are needed. Plus, adding yet another COST to the trucking company or owner operator - without a firm reasoning

(2) The EOBR is a get rich in a hurry gift to the primary company that is making the unit is making a lot of money and on top of that we are forced to pay 40. dollars a month to stay compliant if that is not another way off letting government reach in your pocket to give it to there friends. And the FMCSA wonders why very few owner operators trust them.

1

Figure 4: Excerpt of annotation guidelines