10^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2024, Philadelphia, USA

Making sense of public participation in rulemaking using argument explication

Ankita Gupta, Ethan Zuckerman, Brendan O'Connor

Extended Abstract

Rulemaking, the process by which U.S. federal agencies issue new regulations, involves notifying the public of the proposed rule and soliciting public comments on it, before issuing a final rule. The Administrative Procedures Act 1946 has been interpreted to require the agencies to review and respond to substantive comments [1]. Thus, facilitating comprehension of the large volume of such public comments is crucial for civic decision-making [2]. Prior work has used several content analysis techniques, such as topic modeling [3, 4] and clustering [5], to uncover the main beliefs (or propositions) held by the public. However, comments are often argumentative, where commenters not only state their beliefs but also provide reasons to support them. We report on our current work [6], where we propose a computational method to analyze arguments and apply it to identify and visualize arguments expressed across multiple public comments, thus providing a corpus-level summary.

Task: Public comments on the proposed regulations could be submitted by a wide range of stakeholders (e.g., advocacy groups, and interested individuals) [7]. As a result, some of these comments may lack clear argument structure and reasoning. To analyze such arguments, we propose the task of *argument explication*, which involves making explicit the *structure* and *implicit reasoning* of an argument by decomposing it into the following three core components:

The ***claim*** **(**c**)** is a normative assertion or point of view put forward by the commenter for general acceptance. It is also known as *conclusion* [8–10].

A ***reason*** **(**r_i**)** is a proposition provided by the commenter to convince the audience why they should accept the claim. It is also known as *data* [8], *grounds* [11], and *premise* [9, 10].

The ***warrant*** **(**w_i**)** explains why the claim follows from the reason [8]. It is a missing piece of information, taken for granted and assumed common knowledge by the commenter, yet if it fails to hold, the claim cannot be inferred from the reason. It is similar to *major premise* [9].

Formally, the task input is a textual argument T, and the output is a collection of explication triples, $E = \{\langle c, r_i, w_i \rangle\}\ \forall\ i{=}1$ to N, with the same claim appearing in all triples (see Figure 2).

Method: To explicate an argument, we prompt large language models (LMs) with references to Toulmin's model of argumentation [8]. Toulmin's theory provides a schema to decompose an argument into three core components—*claim*, *grounds* (or *data*), and *warrant*—which map to the components defined in our task; it also has other optional components. Prior work has used this theory to annotate data for supervised model training [12]. In contrast, our approach uses theory references as prompts to steer an LM's response (as per the theory) without requiring any training data. Specifically, we prompt GPT-4 [13] with '`According to Toulmin model,`' which elicits responses with correct mentions of theory *terms* in over 99% cases and generates reasonable *values* (propositions) for each term (see Figure 3). We use this prompt to explicate arguments in two stages. In stage 1, we identify the claim (c) and reasons (r_i) by extracting the values corresponding to *claim* and *grounds* (or *data*) in the LM's response. In stage 2, for each identified claim-reason pair, we generate a warrant (w_i); we input concatenated claim and reason and extract the value for *warrant* from the LM's response.

We evaluate the LM's outputs on prior argumentation datasets [12, 14, 15] and observe that the LM-generated claims and reasons are similar to the gold annotations (Tables 1 and 2). Since a claim-reason pair can admit multiple warrants, we judge the quality of LM-generated warrants via a human evaluation and observe that they are acceptable in 61.7% cases, more often than the gold warrants (45.7%). We conduct further robustness checks, including open-weight LMs, prompting without theory references, and alternative argumentation theories, and observe that across all LMs prompting with references to Toulmin's theory yields better performance [6].

Analyzing public comments: Having established the internal validity of our method, we then apply it to a corpus of 10,000 public comments to the FDA on COVID-19 vaccine approval for children [5]. In particular, we refine clustering-based corpus analysis methods by integrating interconnections among propositional clusters, thus revealing prevalent local argument structures within the discussion. We first obtain propositional clusters by explicating comments (excluding single-sentence comments which are often non-argumentative) and clustering propositions from triples, regardless of their role. We use DP-means clustering [16, 17], which automatically determines the number of clusters based on a distance threshold; we select a threshold of 0.5 based on visual inspection of cluster quality. From 9,187 comments, we obtain 14,137 triples and 308 propositional clusters. To infer interconnections among clusters, we represent a proposition with its cluster ID followed by transforming explication triples (of propositions) into triples of cluster IDs (TIDs), where each TID represents a local argument structure mentioned in one or more comments. Overall, we obtain 6,811 unique TIDs, visualized as a hypergraph,[1] where a propositional cluster is a node and a TID forms a hyperedge.

Interpretive analysis of the corpus based on the hypergraph: We draw several interesting insights from our resultant hypergraph. Among all the TIDs, 1,862 appear in more than one comment, suggesting that people not only share common beliefs but also use similar argument structures to support their beliefs. Figure 1 shows a fragment of the larger argument hypergraph around the most common argument, (c=P1, r=P2, w=P5), which occurs 373 times; it opposes vaccine approval (c=P1) by saying that children have a low risk from the disease (r=P2). Some comments further elaborate on the backing for P2, by citing low mortality rates from COVID-19 among children (P8), obtained by citing data from government websites. On further exploring the local neighborhood of P1, we find two other frequently mentioned reasons: vaccine side-effects (P7) and lack of long-term testing (P3), consistent with findings from studies of social media discussion on vaccines [18], conferring convergent validity to our approach from a different source. Explicitly stating warrants also helps reveal the relationship between distinct parts of the hypergraph.[2] Since we cluster all propositions irrespective of their role in a triple, some clusters include both implicit and explicit propositions. For instance, cluster P5 (vaccines are unnecessary for children) includes propositions implied in some comments, while explicitly stated in others. Thus, such clusters bridge distinct parts of the hypergraph.

Overall, we find corpus visualization as a hypergraph promising direction for future work. Graph visualization (e.g., among concepts, entities) has been long proposed for exploratory corpus analysis [19, 20]. Complementary to these efforts, our approach visualizes 'arguments' at scale, thus aiding the summarization of large volumes of public comments [3]. More broadly, our work demonstrates how generative language models could be used to assist *interpretive* work or content analysis in the computational social sciences.

[1] Unlike a graph, a hypergraph edge—here, a triple $\langle c, r, w \rangle$—can connect more than two nodes.

[2] A claim-reason pair may be linked by several warrants; for visual clarity, we only display the most frequent.

References

[1] Jeffrey Lubbers. State Legislatures. *American Bar Association*, 2006.

[2] Narges Mahyar, Diana V. Nguyen, Maggie Chan, Jiayi Zheng, and Steven P. Dow. The Civic Data Deluge: Understanding the challenges of analyzing large-scale community input. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, page 1171–1181, New York, NY, USA, 2019. Association for Computing Machinery. URL https://doi.org/10.1145/3322276.3322354.

[3] Michael A Livermore, Vladimir Eidelman, and Brian Grom. Computationally assisted regulatory participation. *Notre Dame L. Rev.*, 93:977, 2017.

[4] CDO Council. Implementing federal-wide comment analysis tools. Technical report, 2021. URL https://resources.data.gov/assets/documents/CDOC_Recommendations_Report_Comment_Analysis_FINAL.pdf.

[5] Alexander Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. Natural language decompositions of implicit content enable better text representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13188–13214, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.815.

[6] Ankita Gupta, Ethan Zuckerman, and Brendan O'Connor. Harnessing Toulmin's theory for zero-shot argument explication. *Under Review*, 2024.

[7] Jaime Arguello and Jamie Callan. A bootstrapping approach for identifying stakeholders in public-comment corpora. In *Digital Government Research*, 2007. URL https://api.semanticscholar.org/CorpusID:2711432.

[8] Stephen E. Toulmin. *The Uses of Argument*. Cambridge University Press, 1958.

[9] Douglas Walton. *Argumentation Schemes for Presumptive Reasoning*. Lawrence Erlbaum Associates, Mahwah, New York, 1996. URL https://doi.org/10.4324/9780203811160.

[10] James B. Freeman. *Dialectics and the Macrostructure of Arguments: A Theory of Argument Structure*. De Gruyter Mouton, Berlin, Boston, 1991. URL https://doi.org/10.1515/9783110875843.

[11] S. Toulmin, R.D. Rieke, and A. Janik. *An Introduction to Reasoning*. Macmillan, 1984. URL https://books.google.com/books?id=FTUQAQAAIAAJ.

[12] Ivan Habernal and Iryna Gurevych. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179, April 2017. URL https://aclanthology.org/J17-1004.

[13] OpenAI. GPT-4 Technical Report. *ArXiv*, abs/2303.08774, 2023. URL https://api.semanticscholar.org/CorpusID:257532815.

[14] Andreas Peldszus and Manfred Stede. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon*, volume 2, pages 801–815, 2015.

[15] Maria Becker, Katharina Korfhage, and Anette Frank. Implicit knowledge in argumentative texts: An annotated corpus. *Proceedings of the 12th Conference on Language Resources and Evaluation*, 2020.

[16] Or Dinari and Oren Freifeld. Revisiting DP-Means: Fast scalable algorithms via parallelism and delayed cluster creation. In *Conference on Uncertainty in Artificial Intelligence*, 2022. URL https://api.semanticscholar.org/CorpusID:252306136.

[17] Brian Kulis and Michael I. Jordan. Revisiting k-means: New Algorithms via Bayesian Nonparametrics. In *Proceedings of the 29th International Conference on Machine Learning*, page 148. icml.cc / Omnipress, 2012.

[18] Dominik Wawrzuta, Mariusz Jaworski, Joanna Gotlib, and Mariusz Panczyk. What arguments against COVID-19 vaccines run on Facebook in Poland: Content analysis of comments. *Vaccines*, 9(5):481, 2021.

[19] Abram Handler and Brendan O'Connor. Relational summarization for corpus analysis. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1760–1769, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL https://aclanthology.org/N18-1159.

[20] Tobias Falke and Iryna Gurevych. GraphDocExplore: A framework for the experimental comparison of graph-based document exploration techniques. In Lucia Specia, Matt Post, and Michael Paul, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 19–24, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL https://aclanthology.org/D17-2004.

[21] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*, 2004. URL https://api.semanticscholar.org/CorpusID:964287.

[22] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

[23] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 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, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.741.

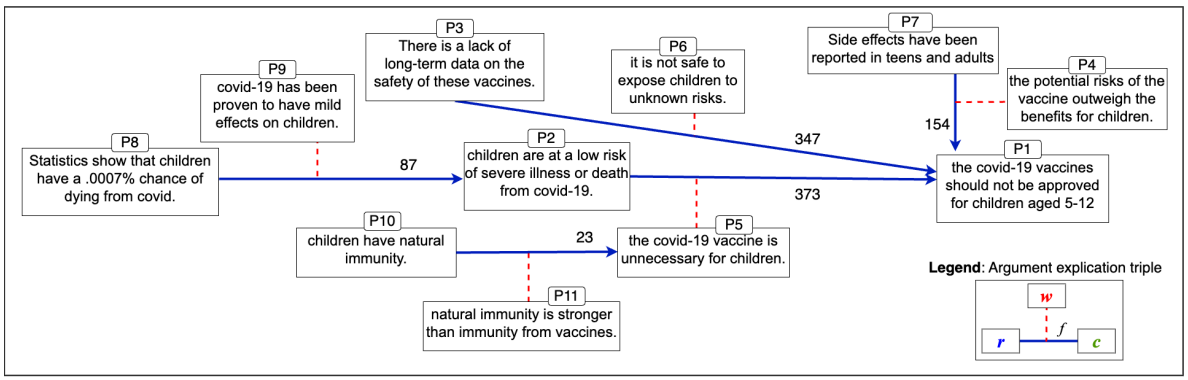

Figure 1: A portion of the corpus-level argument hypergraph we automatically extract from public comments submitted on regulations.gov on whether to approve a COVID-19 vaccine for children. Each node is a cluster of propositions extracted from comments. An argument is a triple of nodes, $\langle$*(c)*laim, *(r)*eason, *(w)*arrant$\rangle$, visualized as *solid blue* and *dotted red* arrows connecting the reason and warrant (r, w) to the claim (c). f is the triple's corpus frequency.

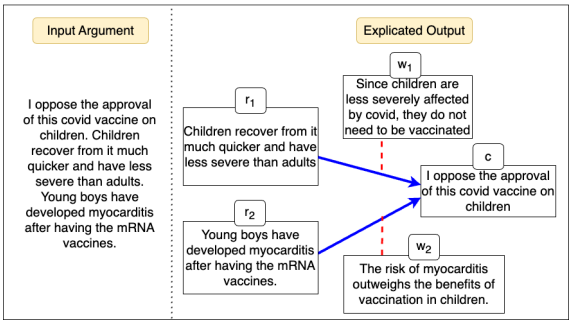

Figure 2: Illustrative example of an input argument decomposed into two explication triples of claim (c), reasons (r_i), and warrants (w_i), visualized as an argument-level hypergraph.

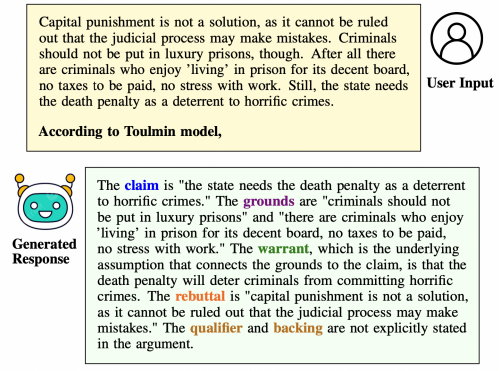

Figure 3: An input argument and an example response obtained by prompting GPT-4 with the `'According to Toulmin model'`. The response correctly mentions *terms* from Toulmin's theory (e.g., claim, grounds) and generates plausible *values* (propositions) for each of these terms.

Prompt	Dataset	BERTScore		Rouge-L	
		Recall	Precision	Recall	Precision
According to Toulmin model,	ARCT	0.99±0.01	0.98±0.01	1.00±0.01	0.98±0.01
	MCT	0.78±0.04	0.79±0.04	0.79±0.05	0.77±0.05
What is the claim of this argument?	ARCT	0.95±0.01	0.91±0.02	0.99±0.01	0.90±0.02
	MCT	0.72±0.03	0.58±0.05	0.69±0.05	0.52±0.06

Table 1: We evaluate GPT-4-generated claims by comparing them with gold claims from two prior argumentation datasets (ARCT [12] with short arguments containing a single claim and a reason, and MCT [14] with longer arguments containing a claim supported by more than one reasons). We measure semantic similarity between gold and generated claim, using ROUGE-L ([21]; n-gram overlap) and BERTScore ([22]; token-level similarity via contextualized word embeddings). We observe that the LM achieves high precision and recall suggesting that the LM-generated claim matches the gold claim. In contrast, when directly asking LM to generate the claim of an argument (i.e., prompting without referring to Toulmin's theory), we observe a low precision, suggesting that LM generates a lot of irrelevant information in addition to the relevant claim.

Prompt	Dataset	Recall	Precision
According to Toulmin model,	ARCT	0.88±0.03	0.87±0.03
	MCT	0.83±0.05	0.86±0.05
What are the reasons provided to support this claim?	ARCT	0.91±0.03	0.93±0.02
	MCT	0.82±0.07	0.75±0.05

Table 2: We also evaluate GPT-4-generated reasons by comparing them with gold reasons from ARCT and MCT. Since the number of gold and generated reasons may differ and the generated reasons may not be strict spans of the input argument but light paraphrases, one-to-one mapping between generated and gold reasons is unknown. To mitigate this issue, we adopt FactScore [23], which measures whether a proposition is supported by a given context. We use FactScore to measure precision (number of generated reasons supported by the concatenated gold reasons) and recall (number of gold reasons supported by concatenated generated reasons). We observe that GPT-4 achieves a high recall and precision on both datasets, suggesting that it can identify all relevant reasons from both short and long arguments without generating irrelevant information. In contrast, when the LM is asked to directly generate reasons, the precision drops on longer arguments from MCT, suggesting that the LMs generate a lot of irrelevant information in addition to the relevant reasons.