Cross-corpora argument analysis using textual entailment

Anonymous authors
Paper under double-blind review

Abstract

We present an *entailment-based clustering method* for conducting cross-corpora argument analysis. Unlike traditional clustering based on semantic similarity, our approach uses natural language inference to cluster textual propositions based on entailment relationships. We then apply our method to 2.34 billion Reddit discussions across three vaccines, namely COVID-19, HPV, and MMR. Our results demonstrate that the entailment-based clustering method better preserves distinct argumentative relations, compared to traditional hierarchical clustering, and uncovers reasoning patterns involved in vaccine hesitancy across multiple different vaccines.

1 Introduction

Facilitating comprehension of the large volume of public opinion is crucial for civic decision-making (Mahyar et al., 2019). Prior work has used several automated content analysis techniques, such as topic modeling (Livermore et al., 2017; Council, 2021) and clustering (Hoyle et al., 2023), to uncover the main beliefs (or propositions) held by the public. However, many times public opinion (e.g., comments, social media posts) can be argumentative, where authors not only state their beliefs but also provide reasons to support them.

In this work, we propose a computational method for cross-corpora argument analysis using *entailment-based clustering*. We then apply it to study arguments shared across three different vaccines (COVID-19, HPV, and MMR) on Reddit. Vaccine hesitancy represents one of the top 10 global health threats declared by World Health Organization (2019), and prior research has observed spillover effects, suggesting that attitudes toward one vaccine can influence perceptions of other vaccines (LaCour & Bell, 2024). Furthermore, Larsson (2020) shows that modern arguments about vaccines tend to mimic common tropes across history, illustrating how similar argumentation patterns persist across different contexts and time periods. Motivated by these studies, we develop a computational method to analyze argumentative structures across different vaccine discussions at scale and identify cross-vaccine reasoning patterns.

2 Dataset

We analyze 2.34 billion Reddit posts across all subreddits from June 2005 to February 2023, using keyword filtering to identify posts related to COVID-19, HPV, and MMR vaccines (Table 1). Since not all posts are argumentative, we develop a classifier using logistic regression with `DistilBERT` embeddings as input features to identify argumentative posts. We train this classifier using posts from subreddits where community rules indicate the expected type of discussions (Table 2). For example, we label posts from `r/unpopularopinion` as argumentative (where users present and defend their opinions) and `r/explainlikeimfive` as non-argumentative (where users mainly ask questions or seek explanations).

3 Entailment-based clustering method

Our method involves two stages: (1) extract and cluster propositions, (2) infer entailment relations among cluster representatives to build an argument graph.

Proposition extraction. We first decompose each post into propositions, which we define as self-contained atomic statements. We prompt `Llama-3.1-8B-Turbo-Instruct` model[1] with a four-shot-prompt adapted from Kamoi et al. (2023) and examples from Peldszus & Stede (2015) (Figure 1). Table 3 shows the number of propositions extracted per vaccine. We then cluster propositions separately for each vaccine using k-means clustering (Lloyd, 1982), with the number of clusters (k) chosen proportionally to the volume of posts for each vaccine. Overall, we obtain $10k$ COVID-19, $1k$ HPV, and $1k$ MMR clusters, with the proposition closest to the cluster centroid as a cluster representative.

Inferring entailment relations. We infer argumentative relations among cluster representatives using the textual entailment task (Jurafsky & Martin, 2025). More specifically, for each cluster representative, we identify all other cluster representatives that entail it (Figure 2), using an NLI classifier.[2] A cluster is then a set of all premises that entail a single claim. Unlike argument retrieval with predetermined queries (Bondarenko et al., 2021; Ein-Dor et al., 2020), we treat each cluster representative as a potential claim and retrieve supporting premises from the corpus. Since a cluster representative can simultaneously serve as a claim (cluster center) and a premise (cluster member), our approach creates a directed graph structure.

Our work relates to prior research on argument visualization and mapping. Earlier approaches focused on manual visualization of individual arguments (Reed, 2001) or collaborative argument mapping tools (Klein, 2012). Closest to our work, Gupta et al. (2024) automate this process by constructing corpus-level argument hypergraphs, drawing edges among propositions when authors explicitly argue in their comment/post that one proposition supports another. In contrast, our entailment-based approach also helps infer relations among propositions that are never explicitly connected by authors, enabling cross-vaccine analysis by identifying similar arguments across vaccine types.

4 Results

Comparison with hierarchical clustering. An alternative approach to obtaining argument graphs is hierarchical clustering (Cohen-Addad et al., 2017), which also organizes propositional clusters into tree-like structures, though based on semantic similarity rather than entailment relations. This method recursively groups propositions into clusters until the inter-cluster distance exceeds a threshold (0.3 in our experiments). We consider three ways to measure inter-cluster distance: single (distance between closest points), complete (distance between farthest points), and average (mean pairwise distance). To compare the two approaches, we applied both hierarchical clustering (HC) and our entailment-based clustering (EC) to our dataset, analyzing clusters with five or more members. For each HC cluster, we computed its Jaccard similarity with all EC clusters and identified the number of HC clusters that have high overlaps (Jaccard ≥ 0.2) with 3 or more different EC clusters. Across all linkage methods, over 30% of HC clusters significantly overlap with three or more EC clusters (Table 4). This observation suggests that semantic similarity often merges distinct argumentative structures that the entailment method preserves.

Qualitative analysis. We next apply our method to analyze cross-vaccine discourse. Figure 2 shows a cluster centered on "I would rather not take the vaccine," revealing shared patterns: both COVID-19 and MMR discussions include religious justifications and illness concerns, while HPV and MMR propositions reference age differently—MMR emphasizes child protection, HPV focuses on avoiding vaccination later in life. Figures 3 and 4 show clusters that share a premise with Figure 2 but support different claims. Figure 3 centers on death concerns, with COVID-19 posts mentioning elderly protection, while MMR focuses on children. All three vaccines share concerns about side effects, with HPV posts listing effects such as infertility and cancer. Finally, Figure 4 highlights a claim related to freedom of choice, including reasons like work/school mandates and constitutional objections.

Overall, we find our entailment-based clustering method as a promising direction for future work, which can help reveal and visualize argumentative patterns across different corpora. More broadly, our work demonstrates how NLP methods could be used to assist interpretive work or aid exploratory content analysis for public opinion research.

[1] Accessed via https://www.together.ai

[2] Using the `BART-large-MNLI` model from https://huggingface.co/facebook/bart-large-mnli

Strong COVID-19	comirnarty, covovax, nuvaxovid, spikevax, vaxzevria
Strong HPV	ceravix, gardasil, 9vhpv, 4vhpv, gardasil-9, gardasil
Strong MMR	m-m-r ii, mmr ii, mmr v, mmrv, proquad, measles-mumps-rubella, measles-mumps-rubella-varicella
Weak COVID-19	2019-ncov, astra zeneca, astrazeneca, biontech, biotech, c 19, c vid, c-19, china virus, chinavirus, chinese vaccine, chinese virus, comirnarty, corona, coronavirus, covid, covid 19, covid-19, covid19, cvd, darona, delta, j and j, j&j, janssen, johnson & johnson, johnson and johnson, kung flu, moderna, mrnav, ncov, novavaxv, omicron, pfizer, pfizer-biontech, rona, rone, roni, rony, roro, sars cov 2, sars-cov-2, sarscov2, the rona, the vid, wuhan virus, j & j, 2019 ncov, 2019ncov
Weak HPV	anal cancer, anal intraepithelial neoplasia, cancer anal, cancer anus, cancer cervical, cancer head neck, cancer penile, cancer penis, cancer vagina, cancer vulva, carcinoma in situ, carcinoma penile, carcinoma penis, carcinoma vagina, carcinoma vulva, cervical, cervical dysplasia, cervical intraepithelial neoplasia, cervical neoplas, genita wart, hpv, hrhpv, hsil, human papillomavirus, laryngeal papillomatosis, lsil, neoplas anal, neoplas anus, neoplas penile, papilloma, papilloma virus, papillomavirus, papillomavirus infections, penile intraepithelial neoplasia, recurrent respiratory papillomatosis, tumor anal, tumor anus, vaginal cancer, vaginal dysplasia, vaginal intraepithelial neoplasia, vulvar intraepithelial neoplasia
Weak MMR	measle, measles, mmr, mump, immunisations, priorix, proquad, rubella, rubellas
Vaccine keywords	booster, boosters, dose, immunisation, immunise, immunised, immunises, immunising, immunization, immunizations, immunize, immunized, immunizes, immunizing, inoculat, inoculate, inoculated, inoculation, jab, jabs, needle, shot, shots, vaccin, vaccinate, vaccinated, vaccinates, vaccinating, vaccination, vaccinations, vaccine, vaccines, vacine, vaxxer, vaxxers, revaccinated

Table 1: List of keywords used to identify posts related to COVID-19, HPV, or MMR vaccines. Strong keywords (typically vaccine names) indicate high likelihood that a post discusses a specific vaccine. Weak keywords, such as disease names or related terms, suggest a post may reference the illness a vaccine prevents but not necessarily the vaccine itself. To identify vaccine-related posts, we first check for strong keywords. If found, the post is included. If not, we check for weak keywords combined with general vaccine-related keywords; posts containing both are also included. Posts meeting neither condition are excluded.

Discourse type	**Subreddits**
Argumentative	r/unpopularopinion
Information seeking	r/explainlikeimfive
Ranting	r/TrueOffMyChest
Storytelling	r/tifu & r/PointlessStories
Symptom checking	r/DiagnoseMe

Table 2: Subreddits used for training the argumentative versus non-argumentative post classifier. Each subreddit's community rules encourage a specific type of discussion, allowing us to label posts as argumentative or non-argumentative based on their source subreddit.

Vaccine	**Extracted propositions**
COVID-19	371,261
HPV	6,896
MMR	13,965

Table 3: Number of propositions extracted from posts of each vaccine.

Base Prompt

Segment the given text into individual unique facts. Each fact should contain only one main verb and use minimal or no clauses. Each fact should be decontextualized, incorporating all the necessary information from the argument to make it self-contained. Ensure all information in the argument is covered and the wording is as unchanged as possible. Please return the original argument if it cannot be decomposed. If the given input is not a complete argument, and thus cannot be decomposed into individual facts, return the original input. Only give facts (or the original argument) and nothing else.

Four Shot Prompt +

Here are some examples:

Text: "Yes, it's annoying and cumbersome to separate your rubbish properly all the time. Three different bin bags stink away in the kitchen and have to be sorted into different wheelie bins. But still Germany produces way too much rubbish and too many resources are lost when what actually should be separated and recycled is burnt. We Berliners should take the chance and become pioneers in waste separation"

Facts: """1. Yes, it's annoying and cumbersome to separate your rubbish properly all the time.

2. Three different bin bags stink away in the kitchen and have to be sorted into different wheelie bins.

3. But still Germany produces way too much rubbish

4. and too many resources are lost when what actually should be separated and recycled is burnt.

5. We Berliners should take the chance and become pioneers in waste separation!"""

Text: "One can hardly move in Friedrichshain or Neukölln these days without permanently scanning the ground for dog dirt. And when bad luck does strike and you step into one of the many 'land mines' you have to painstakingly scrape the remains off your soles. Higher fines are therefore the right measure against negligent, lazy or simply thoughtless dog owners. Of course, first they'd actually need to be caught in the act by public order officers, but once they have to dig into their pockets, their laziness will sure vanish!"

Facts: """1. One can hardly move in Friedrichshain or Neukölln these days without permanently scanning the ground for dog dirt.

2. And when bad luck does strike and you step into one of the many 'land mines' you have to painstakingly scrape the remains off your soles.

3. Higher fines are therefore the right measure against negligent, lazy or simply thoughtless dog owners.

4. Of course, first they'd actually need to be caught in the act by public order officers,

5. but once they have to dig into their pockets, their laziness will sure vanish!"""

Text: "The parkway was opened in 2001 after just under a year of construction and almost two decades of community requests."

Facts: """1. The parkway was opened in 2001.

2. The parkway was opened after just under a year of construction.

3. The parkway was opened after two decades of community requests."""

Text: "Other title changes included Lord Steven Regal and The Nasty Boys winning the World Television Championship and the World Tag Team Championship respectively."

Facts: """1. Lord Steven Regal won the World Television Championship.

2. The Nasty Boys won the World Tag Team Championship."""

Figure 1: Prompt used for decomposing a given post into propositions.

Linkage method	HC clusters overlapping with EC clusters (size $\geq$ 5)	Total count of HC clusters (size $\geq$ 5)	Percent overlap
Single	22	58	38%
Average	51	156	33%
Complete	37	97	38%

Table 4: Comparison of HC cluster overlapping with $\geq$ 3 EC clusters by linkage method.

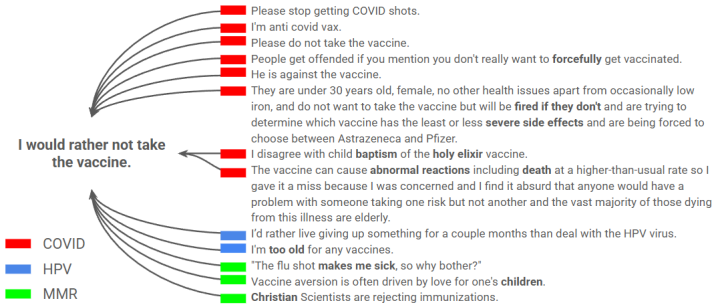

Figure 2: EC cluster centered around "I would rather not take the vaccine."

Figure 3: EC cluster centered around "The vaccines may cause death." The underlined proposition is shared with the EC cluster from Figure 2.

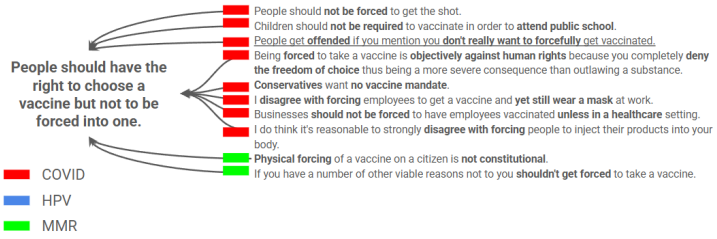

Figure 4: EC cluster centered around "People should have the right to choose a vaccine but not be forced into one." The underlined proposition is shared with the EC cluster from Figure 2.

References

Alexander Bondarenko, Lukas Gienapp, Maik Fröbe, Meriem Beloucif, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. Overview of touché 2021: Argument retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings*, pp. 450–467, Berlin, Heidelberg, 2021. Springer-Verlag. ISBN 978-3-030-85250-4. doi: 10.1007/978-3-030-85251-1_28. URL https://doi.org/10.1007/978-3-030-85251-1_28.

Vincent Cohen-Addad, Varun Kanade, Frederik Mallmann-Trenn, and Claire Mathieu. Hierarchical clustering: Objective functions and algorithms, 2017. URL https://arxiv.org/abs/1704.02147.

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.

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. Corpus wide argument mining - A working solution. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 7683–7691. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6270. URL https://doi.org/10.1609/aaai.v34i05.6270.

Ankita Gupta, Ethan Zuckerman, and Brendan O'Connor. Harnessing toulmin's theory for zero-shot argument explication. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10259–10276, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.552. URL https://aclanthology.org/2024.acl-long.552/.

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*, pp. 13188–13214, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.815.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition, 2025. URL https://web.stanford.edu/~jurafsky/slp3/. Online manuscript released January 12, 2025.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. Wice: Real-world entailment for claims in wikipedia, 2023. URL https://arxiv.org/abs/2303.01432.

Mark Klein. How to harvest collective wisdom on complex problems: An introduction to the mit deliberatorium. MIT Working Paper, 2012. URL https://api.semanticscholar.org/CorpusID:1721610.

Mark LaCour and Zebulon Bell. Attitudes towards COVID-19 vaccines may have "spilled over" to other, unrelated vaccines along party lines in the United States. *Harvard Kennedy School (HKS) Misinformation Review*, 2024. doi: 10.37016/mr-2020-148. URL https://doi.org/10.37016/mr-2020-148.

Paula Larsson. COVID-19 anti-vaxxers use the same arguments from 135 years ago, October 2020. URL https://theconversation.com/covid-19-anti-vaxxers-use-the-same-arguments-from-135-years-ago-145592.

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

S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2): 129–137, 1982. doi: 10.1109/TIT.1982.1056489.

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*, pp. 1171–1181, New York, NY, USA, 2019. Association for Computing Machinery. URL https://doi.org/10.1145/3322276.3322354.

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

Chris Reed. Araucaria: Software for puzzles in argument diagramming and XML. 2001. URL https://api.semanticscholar.org/CorpusID:17256256.

World Health Organization. Ten threats to global health in 2019. https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019, January 2019.