Interactively Uncovering Latent Arguments in Social Media Platforms: A Case Study on the Covid-19 Vaccine Debate

Maria Leonor Pacheco[1,2] **Tunazzina Islam**[3] **Lyle Ungar**[4] **Ming Yin**[3] **Dan Goldwasser**[3]
[1]Microsoft Research [2]University of Colorado Boulder
[3]Purdue University [4]University of Pennsylvania
maria.pacheco@colorado.edu ungar@cis.upenn.edu
{islam32,mingyin,dgoldwas}@purdue.edu

Abstract

Automated methods for analyzing public opinion have grown in popularity with the proliferation of social media. While supervised methods can be very good at classifying text, the dynamic nature of social media discourse results in a moving target for supervised learning. Meanwhile, traditional unsupervised techniques for extracting themes from textual repositories, such as topic models, can result in incorrect outputs that are unusable to domain experts. For this reason, a non-trivial amount of research on social media discourse still relies on manual coding techniques. In this paper, we present an interactive, humans-in-the-loop framework that strikes a balance between unsupervised techniques and manual coding for extracting latent arguments from social media discussions. We use the COVID-19 vaccination debate as a case study, and show that our methodology can be used to obtain a more accurate, interpretable set of arguments when compared to traditional topic models. We do this at a relatively low manual cost, as 3 experts take approximately 2 hours to code close to 100k tweets.

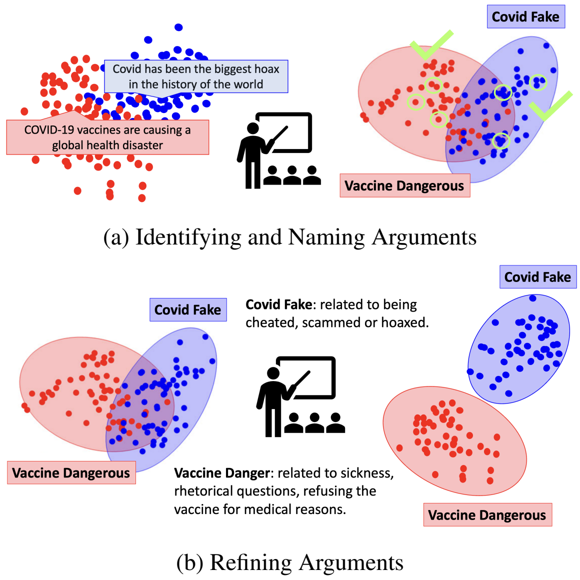


Figure 1: Interactive Framework

1 Introduction

Public opinion plays an important role in the making of policy in pluralistic and democratic societies, as it allows the will of citizens to be heard and accounted for (Smith, 1942; Verba, 1995). As social media has become one of the main outlets for political and civic engagement (Rainie et al., 2012), there is a growing body of work focused on automatically analyzing public opinion on social media. The applications studied include identifying the sentiment towards specific governmental measures (Cortis and Davis, 2019; Wang et al., 2020), detecting and analyzing morally charged statements about current events (Hoover et al., 2020; Pacheco et al., 2022), exploring how ordinary citizens frame political issues (Mendelsohn et al., 2021), and contrasting the stances expressed in social media with public opinion surveys (Joseph et al., 2021). In all of these cases, the variables of interest are well defined, and substantial efforts are dedicated to creating manually annotated resources. In other words, we know which issues, governmental measures, or frames are of interest, and the problems can be framed as supervised learning tasks.

In this paper, we take a step back and tackle the problem of defining the space of relevant variables to analyze public opinion online. Given a widely debated topic, we put the focus on uncovering repeating arguments surrounding discussions on Twitter. Uncovering general themes from collections of unstructured textual resources is a common goal for researchers and practitioners across various disciplines. Unsupervised techniques, like topic models and clustering methods, have been the de-facto NLP solution to this problem (Blei et al., 2003; Boyd-Graber et al., 2017; Sia et al., 2020). While widely used, these methods often produce topics that are clearly incorrect to domain experts

Proceedings of the Fourth Workshop on Data Science with Human-in-the-Loop (Language Advances), pages 94 - 111
December 8, 2022 ©2022 Association for Computational Linguistics

(Mimno et al., 2011). For this reason, and despite the popularity of contemporary NLP techniques, a non-trivial number of recent studies on social media discourse rely on manual, qualitative coding methods instead (Valle et al., 2020; Nguyen et al., 2021; Hagen et al., 2022).

In this work, we strike a balance between unsupervised NLP techniques and manual coding by adopting a humans-in-the-loop approach. We use the Twitter debate surrounding the COVID-19 vaccine as a case study, and present an interactive framework to discover and define the space of arguments frequently cited as reasons to refuse or accept the vaccine. Our framework is designed to address two main challenges: 1) Given a large collection of tweets, how can experts effectively explore the data and identify a set of repeating arguments, and 2) Given a known space of high-level arguments, how can experts create and refine a representation that improves the mapping from tweets to arguments. For example, in Fig. 1a we can observe an initial clustering of arguments provided by the system. Our initial goal is to obtain a resulting set of named arguments with high-quality examples as identified by the experts. Similarly, In Fig. 1b we can observe two overlapping arguments. Our next goal is to source explanations from the experts that result in a better mapping from tweets to arguments. This can be achieved by expanding and refining the theme representation according to the explanations provided. The expected result is a comprehensive set of high-level arguments that explain the discussion about the COVID-19 vaccine, and a partial mapping from tweets to their most likely argument.

To tackle our goals, we design and implement a simple protocol that allows groups of experts to work together towards this goal, and introduce an interactive tool equipped with operations to facilitate the discovery and refinement of arguments in large language resources. Our work is related to interactive systems that leverage clustering techniques to help users discover relevant topics (Bernstein et al., 2010), systems that exploit visualization techniques to label data interactively (Bernard et al., 2017, 2018; Vajiac et al., 2022), and human-in-the-loop topic modeling approaches that let users refine discovered topics (Hu et al., 2011; Lund et al., 2017; Smith et al., 2018). The main differences between these systems and our work are that: 1) we do not assume a known space of relevant labels, 2) we let experts drive and influence the topic discovery procedure in addition to supporting the exploration, and 3) we support open-ended feedback from experts in natural language.

Our experiments show that our framework can be used to uncover a set of arguments that cover a large portion of the discussion about the COVID-19 vaccine on Twitter, and that the resulting mapping from tweets to argument is fairly accurate with respect to human judgements. Additionally, we use the dataset of tweets about the COVID-19 vaccine released by Pacheco et al. (2022), which is annotated for vaccination stance and morality frames, and show that the high-level arguments obtained using our methodology have higher correlations with vaccination stance and moral sentiments than topics obtained using traditional topic models.

2 Interactive Framework

We propose a simple protocol that combines NLP techniques, interactive interfaces and qualitative methods to assist experts in characterizing large tweet repositories about the COVID-19 vaccine. Our protocol takes a large repository of tweets and automatically proposes an initial partition of the data, such that tweets that are thematically similar are clustered together. We provide experts with an interactive interface equipped with a set of exploratory operations that allows them to evaluate the quality of the discovered clusters, as well as to further explore and partition the space by inspecting individual examples, finding similar tweets, and using open text queries. As they interact with the data through interface, a group of experts work together following an inductive thematic analysis approach to identify and code the patterns that emerge within the partitions (Braun and Clarke, 2012). Next, they group the identified patterns into general arguments, and instantiate them using the interface. Although intuitively we could expect a single cluster to result in a single argument, note that this is not enforced. Experts maintain full freedom as to how many arguments they instantiate, if any. Once an argument is created, experts are provided with a set of operations to explain the argument using natural language, select good example tweets, or write down additional examples. At any point during the process, experts can toggle a procedure that assigns tweets to arguments.

Operations	Description
Finding Clusters	Experts can find clusters in the space of unassigned tweets. To do this, we run a clustering algorithm using the tweet representations described in Sec. ??. We support the K-means (Jin and Han, 2010) and Hierarchical Density-Based Clustering (McInnes et al., 2017) algorithms. For all results presented in this paper, we use the K-means algorithm.
Text-based Queries	Experts can type any query in natural language and find tweets that are close to the query in the embedding space.
Finding Similar Tweets	Experts have the ability to select each tweet and find other examples that are close in the embedding space.

Table 1: Discovery Operations

2.1 Interactive Tool

To support our interactive framework, we developed a tool for experts to interact with a large number of tweets. The tool is a simple GUI equipped with a finite set of exploratory and intervention operations. *Exploratory operations* allow experts to discover clusters of tweets and further explore and partition the space, and to evaluate the quality of the discovered clusters and the grounded statements. *Intervention operations* allow experts to name the discovered patterns, as well as to provide examples and judgements to improve the quality of the initial partitions.

Representing Tweets and Arguments: We represent tweets using their Sentence BERT embedding (Reimers and Gurevych, 2019). We represent arguments using a handful of explanatory phrases and a small set of examples, and calculate their SBERT embeddings. Note that our tool is agnostic of the representation used, as the underlying embedding objective can be easily replaced.

Exploratory Operations: These operations allow experts to inspect the current state of the system, both to evaluate the quality of the tweet-argument mappings, as well as to explore the data and discover new emerging arguments. We divide exploratory operations in two types: discovery operations and quality assurance operations. *Discovery operations* allow users to explore the space of tweets and get a sense of what arguments emerge in the data. We enumerate them in Tab. 1. *Quality assurance operations* allow users to evaluate the quality of the discovered clusters and the grounded tweets. We enumerate them in Tab. 2.

Intervention Operations: These operations allow experts to introduce knowledge into the system to improve the discovery and grounding of emerging arguments. We enumerate them in Tab. 3.

Cropped screenshots demonstrating all of these operations can be observed in Appendices, A.1, A.2 and A.3. Additionally, we include screenshots of the full view of all pages in our GUI in Appendix A.4.

Operations	Description
Listing Arguments and Examples	Experts can browse the current list of arguments and their grounded examples. Examples are ranked in order of "goodness", corresponding to the similarity in the embedding space to the argument representation. Examples are listed from closest to most distant, or from most distant to closest.
Visualizing Local Explanations	Experts can visualize aggregated statistics and explanations for each of the grounded arguments. To obtain these explanations, we aggregate all instances that have been identified as being associated with a theme. Explanations include wordclouds, frequent entities and their sentiments, and graphs of feature distributions.
Visualizing Global Explanations	Experts can visualize aggregated statistics and explanations for the global state of the system. To do this, we aggregate all instances in the database. Explanations include argument distribution, coverage statistics, and 2-Dimensional t-sne plots (van der Maaten and Hinton, 2008).

Table 2: Quality Assurance Operations

Operations	Description
Adding and Removing Arguments	Experts can create and remove arguments. The only requirement for creating a new argument is to give it a unique name. Similarly, arguments can be removed at any point. If any instances are assigned to an argument being removed, they will be assigned to the *"Unknown"* argument.
Adding and Removing Examples	Experts can assign "good" and "bad" examples to existing arguments. Good examples are instances that characterize the named argument. Bad examples are instances that could have similar wording to a good example, but that have different meaning. Experts can add examples in two ways: they can mark grounded tweets as "good" or "bad", or they can directly contribute example phrases.

Table 3: Intervention Operations

Argument Grounding: At any point during interaction, experts can toggle a procedure that assigns tweets to arguments. We use a simple distance-based approach for this purpose. To measure the closeness between a tweet and an argument, we compute the cosine distance between the tweet and *all* of the explanatory phrases and examples for the argument, and take the minimum distance score among them. Before this operation is called for the first time, all tweets belong to *unnamed clusters*. In other words, they remain unassigned. Once this operation is called, we assign tweets to their closest argument *if and only if* the newly computed distance is less than or equal to the distance to its previous assignment. Previous assignments can correspond either to different arguments, or to the unnamed space. Note that this way, some tweets can remain unassigned.

3 Case Study

As a case study, we look at tweets written about the COVID-19 vaccine on Twitter. We collected a corpus of 85,000 tweets that mentioned the vaccine. To avoid repetitions, we filter out all retweets ahead of time. The collected tweets are uniformly distributed between January and October, 2021. All tweets in our corpus are written in English, and were posted by users located in the United States. Our main goal is to use the framework introduced in Sec. 2 to identify repeating themes in this corpus, and construct a set of high-level arguments that are frequently used to justify stances on the vaccine.

For our study, we recruited six experts in Natural Language Processing and Computational Social Science, four male and two female, within the ages of 25 and 45. The experts included graduate students, postdoctoral researchers and faculty. To evaluate the different components of our framework, we performed a two-stage analysis. In the first stage, we simplify the problem and assume that we have an initial, known set of high-level arguments, and let three of the experts focus on the challenge of interactively refining the arguments and grounding them in the large Twitter corpus. In the second stage, we remove this assumption and have the remaining three experts discover the space of relevant arguments from scratch. Below, we present each of these scenarios in detail and perform both qualitative and quantitative evaluations to assess the outcome of the interaction.

3.1 Stage 1: Mapping Tweets to Arguments

In this stage, we assume that we know what is the set of relevant arguments, and our main goal is to improve the mapping between tweets and arguments. We build on previous work on health informatics studying the arguments made by Twitter users in Poland when discussing the COVID-19 vaccine (Wawrzuta et al., 2021). This work introduces a code-book of 13 main arguments defined using short phrases in natural language (Tab. 4).

We start by mapping the collection of 85k tweets to the Wawrzuta et al. (2021) arguments using the distance between their SBERT embeddings. Then, we let the experts interact with the system following the protocol outlined in Sec. 2. Below, we outline the interactive sessions performed by the three experts in detail.

Interactive Sessions: The three experts started by looking at the global visualizations. Then, they inspected the arguments one by one, looking at the local explanations and the 10 closest and 10 furthest tweets from each argument. Next, they were involved in a discussion phase to identify arguments that were present in the data, but not covered by the Wawrzuta et al. (2021) set, as well as the argumentation patterns that the system failed to identify for each of the arguments. This process was done in two one-hour sessions.

1	Lack of trust in the government
2	The vaccine will be dangerous to health
3	The COVID-19 vaccine disease does not exist
4	I do not want to be vaccinated because I have freedom of choice
5	The vaccine was created for the profit of pharmaceutical companies
6	Natural methods of protection are better than the vaccine
7	The vaccine does not work
8	The vaccine is not properly tested, it was developed too quickly
9	No one is responsible for the potential side effects of the vaccine
10	Mentioning past development of the swine-flu vaccine
11	The vaccine existed before the epidemic, there is too much resistance
12	Conspiracy theories, hidden vaccine effects (e.g. chips)
13	Positive attitude towards the vaccine

Table 4: 13 Arguments Proposed by Wawrzuta et al. (2021)

PRO VAX	government distrust, vaccine dangerous, covid fake, vaccine oppression, pharma bad, natural immunity effective, vaccine against religion, vaccine does not work, vaccine not tested, bill gates' micro chip, vaccine tested on dogs, vaccine has fetal tissue, vaccine makes you sterile
ANTI VAX	government trust, vaccine safe, covid real, vaccine not oppression, pharma good, natural immunity ineffective, vaccine not against religion, vaccine works, vaccine tested

Table 5: Resulting Arguments

Initially, the experts focused on adding missing arguments and removing arguments that were not frequently referenced in the data. For example, they noticed that the Wawrzuta et al. (2021) set contained mostly anti-vaccine arguments, and added the positive counterpart for each argument (e.g. *"The vaccine is dangerous"* $\Rightarrow$ *"The vaccine is safe"*). In addition to this, they observed and added new arguments such as *"The vaccine is against my religion"*, and separated *"Conspiracy theories and hidden effects"* into sub-arguments related to particular conspiracy theories, such as *"The vaccine contains fetal tissue"*, and *"The vaccine makes you sterile"*. They also removed infrequent arguments, such as *The swine-flu vaccine*, and came up with shorter names/identifiers for each one of the arguments. The resulting set of arguments can be observed in Tab. 5.

Next, the experts turned their attention to identifying the argumentative patterns that were not being captured by the given argument descriptions. They did this by looking at assignments to other arguments that appeared to be a mismatch, as well as inspecting low confidence assignments. Here, the coders followed a qualitative thematic analysis approach to code relevant patterns. For example, in the case of *"vaccine oppression"*, the experts noted that tweets that included legal terms were not being captured, as well as sarcastic expressions, and tweets that had explicit mentions to discrimination and oppression. They followed this process for every argument, and coded the missing argumentative patterns. Then, each expert contributed a set of 2-5 examples for each argument. In Appendix A.5 we include tables enumerating the full list of coded patterns and contributed phrases.

Evaluation: To evaluate the performance of our tweet to argument mapping in the dataset of 85k unlabeled tweets, we sorted the tweets according to their semantic distance to their assigned arguments, computed the three quartiles, and sampled a set of 12 tweets per argument such that 3 tweets are randomly sampled from each interval. Then, we manually annotated whether the mapping was correct or not. We did this both for the initial mapping, before any interaction, and for the resulting mapping, after interaction. This resulted in 156 tweets and 264 tweets, respectively.

To evaluate the performance at different degrees of semantic distance to the argument embedding, we perform the evaluation at each quartile. Results for the first quartile (Q_1) correspond to the 25% closest examples. For the second quartile (Q_2), they correspond to the 50% closest examples, and for the third quartile (Q_3), to the 75% closest examples. Intuitively, we expect better average performance the lower the distance is between the tweets and the argument. Results are outlined in Tab. 6. While both before and after interaction we have comparable performance for the semantically closest tweets, performance degrades faster using the initial set of arguments. This result makes sense, given that for the Wawrzuta et al. (2021) set we are only relying on one short phrase to represent arguments. The positive impact of refining arguments interactively by enriching the argument representation is clear.

Given that we can characterize arguments as the reasons cited by people to accept or refuse the COVID-19 vaccine, we consider assignments to be better if they are more cohesive (e.g. if they are more strongly correlated with vaccination stance). To evaluate this, we perform a correlation test between the identified arguments and the stance expressed in the tweet (i.e. pro or anti-vaccine). To do this, we use the set of 750 tweets annotated for stance and moral foundations released by Pacheco et al. (2022). We calculate the Pearson correlation matrices and present them in Fig. 2. We compare the arguments obtained interactively with the seed set of manual arguments (Wawrzuta et al., 2021), and with a set of topics extracted using Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a generative, unsupervised topic modeling technique that allows a set of textual instances to be explained by unobserved groups of words that explain their similarity. We can observe that our refined arguments (Fig. 2d) have higher, more accurate correlations with vaccination stances and than both the original set of arguments (Fig. 2b) and the derived LDA topics (Fig. 2a, 2c). For example, we find that in the initial arguments baseline, both *"Vaccine Doesn't Work"* and *"Covid Fake"* have a high correlation with the *"pro-vax"* stance, which is opposite from what would be expected. This behavior is corrected after interaction.

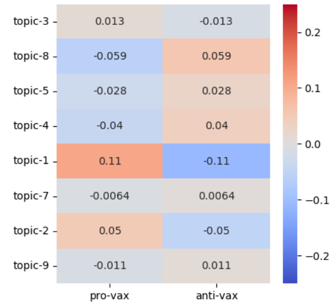
Iter.	# Args	Q_1	Q_2	Q_3	All
Before Interaction	13	89.36	73.81	60.87	52.05
After Interaction	22	88.52	84.87	81.98	80.27

Table 6: Argument F1 w.r.t Human Judgements

3.2 Stage 2: Uncovering Latent Arguments

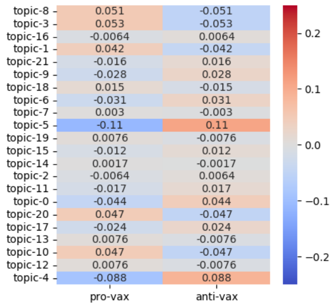
In this stage, we address the challenge of discovering the space of arguments that emerge from our corpus of 85k tweets about the COVID-19 vaccine. Unlike the scenario presented before, we do not assume any prior knowledge, and we make no assumptions about the number of relevant arguments or what they ought to be. Our main goal is to let the three experts leverage our interactive framework to find a set of relevant arguments, as well as a final mapping from tweets to arguments.

The main challenge in this stage is to obtain a set of arguments that accounts for as many tweets as possible, while maintaining the cohesiveness of the partitions and the accuracy of tweet to argument assignments. Below, we explain the interactive process in detail and present an evaluation of the results obtained.

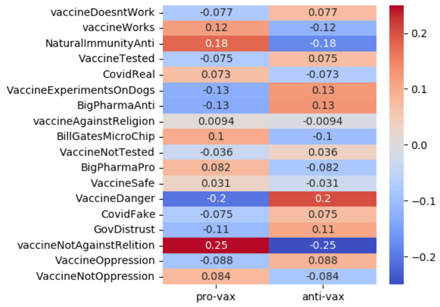


(a) **Baseline**: 10 LDA Topics

(b) **Baseline**: Manual (Wawrzuta et al., 2021)



(c) **Baseline**: 20 LDA Topics



(d) **Ours**: After Interaction

Figure 2: Correlations between arguments and stance

Interactive Sessions: To initialize the system, the experts started by using the clustering operation to find 10 initial clusters of roughly the same size. First, they examined the clusters one by one, looking at the examples closest to the centroid. This was followed by a discussion phase, in which the experts coded the argumentative patterns observed. If one or more cohesive patterns were identified, the experts created a new argument, named it, and marked a set of good example tweets that helped to characterize the named argument. In Appendix A.6 we include a table showing each initial cluster, the argumentative patterns identified, and the named arguments chosen by the experts during the discussion phase. When a pattern was not obvious, the experts explored similar instances to the different tweets found. Whenever the similarity search resulted in a new pattern, the experts coded it and created a new argument.

Next, the experts looked at the local argument explanations and repeated a process similar to the first stage, by enhancing each argument with additional example phrases. Note that each argument already contained a small set of representative tweets, which were marked as "good" in the previous step. Finally, the experts toggled the nearest neighbors operation to map tweets to arguments.

We performed two iterations of this process. In the second iteration, the experts used the clustering operation again over the set of tweets that remained unassigned to existing arguments. Then, they repeated the full process a second time to uncover new arguments. The full table outlining the clusters, coded patterns and resulting arguments for the second iteration are also provided in Appendix A.6.

Evaluation: As in the previous stage, we evaluate the performance of our tweet to argument mapping by sampling a random set of 12 tweets per argument after each iteration of interaction, 3 from each interval. This resulted in a set of 108 tweets for iteration 1 and 192 tweets for iteration 2. Then, we manually annotated whether the mapping was correct or not. To evaluate the performance at different degrees of semantic distance to the argument embedding, we perform the evaluation at each quartile. Results are outlined in Tab. 7.

As expected, we obtained higher F1 scores for tweets that are the closest to the arguments in the embedding space. In addition to the F1 scores, we also look at the percentage of tweets that are covered by the set of arguments uncovered by the

Iter.	# Args	Coverage	Q_1	Q_2	Q_3	All
1	9	9.3%	89.80	87.50	87.50	85.71
2	16	22.9%	90.91	87.06	84.34	77.32

Table 7: Argument F1 w.r.t Human Judgements

experts after each iteration. We remind the reader that we do not enforce that all tweets need to be mapped to arguments, and therefore some tweets remain unassigned. There is a degradation in performance after subsequent iterations, as we increase both the number of arguments and the amount of tweets mapped. However, we find that the gain in coverage is proportionally greater than the drop in performance (x2.5 vs. x1.1). The intuition behind performing subsequent iterations is that we force the system to look at new, previously ignored partitions of the data to find new arguments. In future work, we would like to study how to estimate the optimal number of iterations, as well as when to decide to stop exploring the unassigned space.

4 Conclusions

We presented an initial step towards an interactive, humans-in-the-loop framework for uncovering latent arguments in social media discourse. We implemented a simple protocol that allows groups of experts to work together efficiently to create a comprehensive code-book of high-level arguments, and developed a GUI with a set of computational operations to streamline their coding process. We used the COVID-19 vaccine debate as a case study, and showed that by applying subsequent runs of our methodology, experts can obtain a comprehensive set of arguments that account for a reasonable slice of the data without sacrificing performance. Additionally, we showed that our resulting set of arguments is cleaner and more explainable than themes obtained with topic modeling approaches.

References

Jurgen Bernard, Marco Hutter, Matthias Zeppelzauer, Dieter Fellner, and Michael Sedlmair. 2018. Comparing visual-interactive labeling with active learning: An experimental study. *IEEE transactions on visualization and computer graphics*.

Jürgen Bernard, Matthias Zeppelzauer, Michael Sedlmair, and Wolfgang Aigner. 2017. A Unified Process for Visual-Interactive Labeling. In *EuroVis Workshop on Visual Analytics (EuroVA)*. The Eurographics Association.

Michael S. Bernstein, Bongwon Suh, Lichan Hong, Jilin Chen, Sanjay Kairam, and Ed H. Chi. 2010. Eddi: interactive topic-based browsing of social status streams. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, UIST '10, pages 303–312, New York, NY, USA. ACM.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Jordan Boyd-Graber, Yuening Hu, and David Minmo. 2017. *Applications of Topic Models*.

Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.*, pages 57–71.

Keith Cortis and Brian Davis. 2019. A social opinion gold standard for the Malta government budget 2018. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 364–369, Hong Kong, China. Association for Computational Linguistics.

Loni Hagen, Ashley Fox, Heather O'Leary, DeAndre Dyson, Kimberly Walker, Cecile A Lengacher, and Raquel Hernandez. 2022. The role of influential actors in fostering the polarized covid-19 vaccine discourse on twitter: Mixed methods of machine learning and inductive coding. *JMIR Infodemiology*, 2(1):e34231.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Yuening Hu, Jordan Boyd-Graber, and Brianna Satinoff. 2011. Interactive topic modeling. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 248–257, Portland, Oregon, USA. Association for Computational Linguistics.

Xin Jin and Jiawei Han. 2010. *K-Means Clustering*, pages 563–564. Springer US, Boston, MA.

Kenneth Joseph, Sarah Shugars, Ryan Gallagher, Jon Green, Alexi Quintana Mathé, Zijian An, and David Lazer. 2021. (mis)alignment between stance expressed in social media data and public opinion surveys. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 312–324, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jeffrey Lund, Connor Cook, Kevin Seppi, and Jordan Boyd-Graber. 2017. Tandem anchoring: a multiword anchor approach for interactive topic modeling. In

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 896–905, Vancouver, Canada. Association for Computational Linguistics.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. Modeling framing in immigration discourse on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263, Online. Association for Computational Linguistics.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Thu T. Nguyen, Shaniece Criss, Eli K. Michaels, Rebekah I. Cross, Jackson S. Michaels, Pallavi Dwivedi, Dina Huang, Erica Hsu, Krishay Mukhija, Leah H. Nguyen, Isha Yardi, Amani M. Allen, Quynh C. Nguyen, and Gilbert C. Gee. 2021. Progress and push-back: How the killings of ahmaud arbery, breonna taylor, and george floyd impacted public discourse on race and racism on twitter. *SSM - Population Health*, 15:100922.

Maria Leonor Pacheco, Tunazzina Islam, Monal Mahajan, Andrey Shor, Ming Yin, Lyle Ungar, and Dan Goldwasser. 2022. A holistic framework for analyzing the COVID-19 vaccine debate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5821–5839, Seattle, United States. Association for Computational Linguistics.

Lee Rainie, Aaron Smith, Kay Lehman Schlozman, Henry E. Brady, and Sidney Verba. 2012. Social media and political engagement. *Pew Research Center Reports*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online. Association for Computational Linguistics.

Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2018. Closing the loop: User-centered design and evaluation of a human-in-the-loop topic modeling system. In *23rd International Conference on Intelligent User Interfaces*, IUI '18, page 293–304, New York, NY, USA. Association for Computing Machinery.

Charles W. Smith. 1942. Democracy through public opinion. by harold d. lasswell. (menasha, wisconsin: George banta publishing company. 1941. pp. 176. 1.50.). *American Political Science Review*, 36(2):348–349.

Catalina Vajiac, Duen Horng Chau, Andreas Olligschlaeger, Rebecca Mackenzie, Pratheeksha Nair, Meng-Chieh Lee, Yifei Li, Namyong Park, Reihaneh Rabbany, and Christos Faloutsos. 2022. Trafficvis: Visualizing organized activity and spatio-temporal patterns for detecting and labeling human trafficking. *IEEE transactions on visualization and computer graphics*.

Marc Esteve Del Valle, Rimmert Sijtsma, Hanne Stegeman, and Rosa Borge. 2020. Online deliberation and the public sphere: Developing a coding manual to assess deliberation in twitter political networks. *Javnost - The Public*, 27(3):211–229.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Sidney Verba. 1995. The citizen as respondent: Survey research and american democracy. *American Political Science Review*, March.

Shihan Wang, Marijn Schraagen, Erik Tjong Kim Sang, and Mehdi Dastani. 2020. Public sentiment on governmental COVID-19 measures in Dutch social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Dominik Wawrzuta, Mariusz Jaworski, Joanna Gotlib, and Mariusz Panczyk. 2021. What arguments against covid-19 vaccines run on facebook in poland: Content analysis of comments. *Vaccines*, 9(5):481.

A Appendix

In this section, we include cropped screenshots of the different operations outlined in Section 2.1, as well as full screenshots of all the views of the GUI. Additionally, we include tables with the full results of the qualitative thematic analysis procedures.

A.1 Discovery Operations

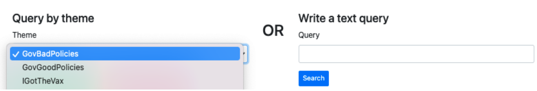
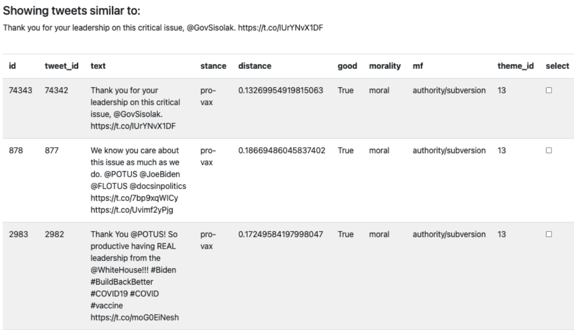


Figure 3: Text-based Queries

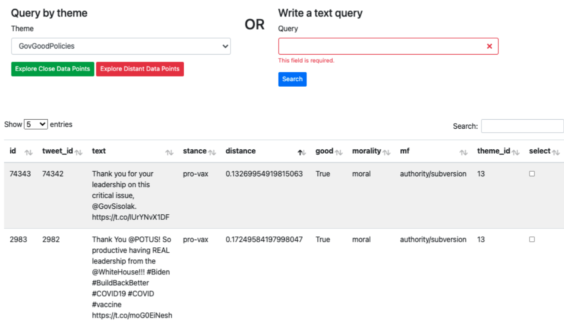


Showing tweets similar to:
Thank you for your leadership on this critical issue, @GovSisolak. https://t.co/lUrYNvX1DF

id	tweet_id	text	stance	distance	good	morality	mf	theme_id	select
74343	74342	Thank you for your leadership on this critical issue, @GovSisolak. https://t.co/lUrYNvX1DF	pro-vax	0.13269954919815063	True	moral	authority/subversion	13	☐
878	877	We know you care about this issue as much as we do. @POTUS @JoeBiden @FLOTUS @docsinpolitics https://t.co/7bp9xqWICy https://t.co/Uvimf2yPjg	pro-vax	0.18669486045837402	True	moral	authority/subversion	13	☐
2983	2982	Thank You @POTUS! So productive having REAL leadership from the @WhiteHouse!!! #Biden #BuildBackBetter #COVID19 #COVID #vaccine https://t.co/moG0EiNesh	pro-vax	0.17249584197998047	True	moral	authority/subversion	13	☐

Figure 4: Finding Similar Tweets

A.2 Quality Assurance Operations



id	tweet_id	text	stance	distance	good	morality	mf	theme_id	select
74343	74342	Thank you for your leadership on this critical issue, @GovSisolak. https://t.co/lUrYNvX1DF	pro-vax	0.13269954919815063	True	moral	authority/subversion	13	☐
2983	2982	Thank You @POTUS! So productive having REAL leadership from the @WhiteHouse!!! #Biden #BuildBackBetter #COVID19 #COVID #vaccine https://t.co/moG0EiNesh	pro-vax	0.17249584197998047	True	moral	authority/subversion	13	☐

Figure 5: Listing Arguments and Examples

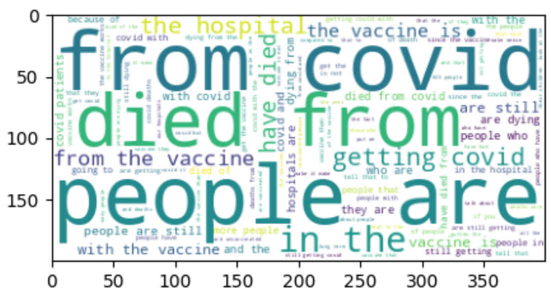
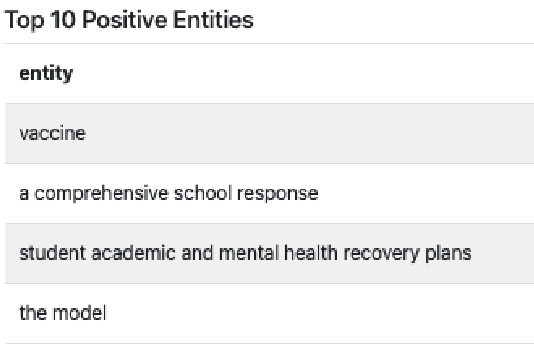


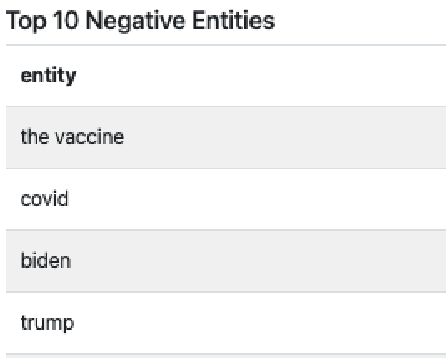
Figure 6: Visualizing Local Explanations: Word Cloud Example for *The Vaccine Doesn't Work*



Top 10 Positive Entities

entity
vaccine
a comprehensive school response
student academic and mental health recovery plans
the model

(a) Top Positive Entities

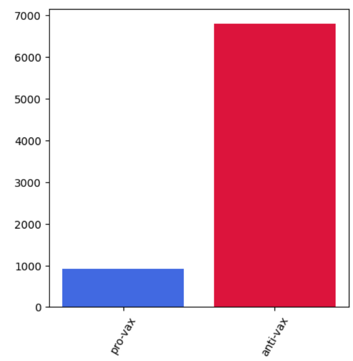


Top 10 Negative Entities

entity
the vaccine
covid
biden
trump

(b) Top Negative Entities

Figure 7: Visualizing Local Explanations: Most Frequent Positive and Negative Entities for *Bad Governmental Policies*



(a) Stance

Figure 8: Visualizing Local Explanations: Attribute Distribution for *The Vaccine Doesn't Work*. Note that attributes can be predicted using external resources. In this case, we predicted stance using a classifier trained on hashtags, as described in (Pacheco et al., 2022).

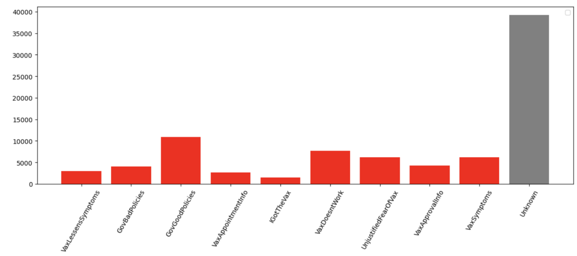


Figure 9: Visualizing Global Explanations: Argument Distribution

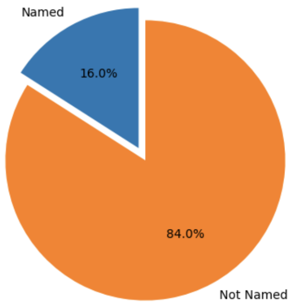


Figure 10: Visualizing Global Explanations: Coverage

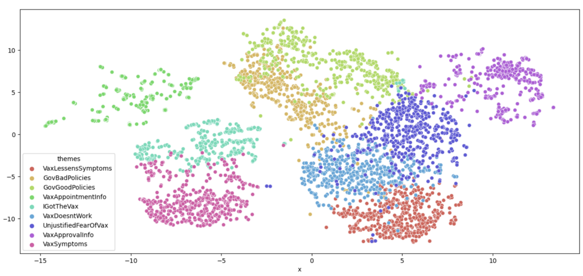


Figure 11: Visualizing Global Explanations: 2D t-SNE

A.3 Intervention Operations

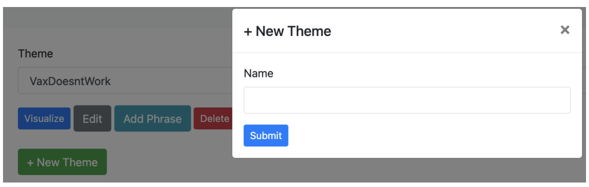


Figure 12: Adding New Themes

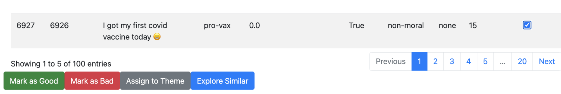


Figure 13: Marking Instances as *Good*

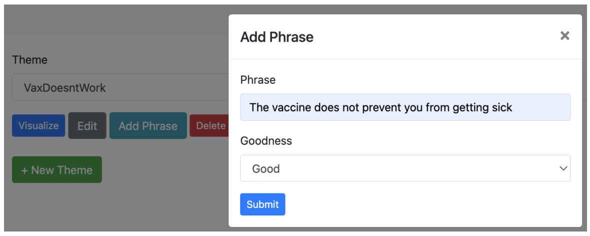


Figure 14: Adding *Good* Examples

A.4 Full Screenshots

Full screenshots of the page views of our GUI can be seen in Figures 15, 16, 17, 18, 19, 20, 21 and 22

A.5 Stage 1: Coded Patterns and Contributed Phrases

Coded patterns for each argument can be observed in Tab. 8. The resulting list of added and removed arguments, as well as their contributed phrases can be observed in Tabs. 9 and 10.

A.6 Stage 2: Argumentative Codes and Resulting Arguments

The clusters for the first iteration of interaction, the coded argumentative patterns and the resulting arguments can be observed in Tab. 11. The same content for the second iteration of interaction can be observed in Tab. 12.

Argument	Argumentative Patterns
GovDistrust	Add phrases with strong word for distrust “Good at being bad” Explicit negations
GovTrust	Hedging phrases (sort-of trust)
VaxDanger	Closer connection between vaccine words and danger words (related to sickness, bad effects) Explit negations Rhetorical questions Refusing the vaccine for medical reasons
VaxSafe	Explicit mentions of safety Explicit negations
CovidFake	Stronger relevant negative words (fake, scam, hoax) Explicit negations
CovidReal	Trust the science References to Covid hospitalization on the rise, explicit mentions of hospitals Explicit negations
VaxOppression	Legal language Explicit mentions of discrimination and oppression Sarcasm
VaxNotOppression	Justifying mandates Freedom to be protected Criticizing others using “you/people” language, focus freedom on me/my/I
BigPharmaAnti	Stronger words against pharmaceutical companies (corrupt, evil) Not accountable / irresponsible past behavior Mentions of negative side-effect of other products (cancer)
BigPharmaPro	Trust science/research and vaccine development process Language about intent, the vaccine was created to do something good, explicit names of companies
NaturalImmunityPro	The vaccine is not enough Explicit mentions to population immunity, herd immunity and antibodies
NaturalImmunityAnti	Emphasis on global look, collective entities, society Natural immunity characterized as dangerous or not effective Mentions of experts and trusting science
VaxAgainstReligion	I put it in god hands (god is deciding) Treating pro-vax as another religion
VaxNotAgainstReligion	“Religious” in quotes Bugus exemptions “Where is your faith” Call to action: get tested/get vaccinated/put a mask on (mentions of compassion) No religion ask members to refuse vaccine
VaxDoesntWork	Reference to “magic vaccine” “Never developed”, “doesn’t work” Questions: why are deaths high? Why is corona not going away? Why are vaccinated people dying?
VaxWorks	“ask a doctor”, consult with an expert Research on the vaccine is good/has been going on for a long time Capture differences, e.g. “good trials” vs. rushed ones.
VaxNotTested	Language suggesting “rushed through trials” and “experimental vaccine”
VaxTested	trust the research and development process Testing can be confused with covid-test, use other language.

Table 8: Coded Argumentative Patterns for Stage 1

Arguments	Contributed Phrases
GovDistrust	**"lack of trust in the government"**, "Fuck the government", "The government is a total failure", "Never trust the government", "Biden is a failure", "Biden lied people die", "The government and Fauci have been dishonest", "The government always lies", "The government has a strong record of screwing things up", "The government is good at screwing things up", "The government is screwing things up", "The government is lying", "The government only cares about money", "The government doesn't work logically", "Do not trust the government", "The government doesn't care about people's health", "The government won't tell you the truth about the vaccine"
VaxDanger	**"the vaccine will be dangerous to health"**, "Covid vaccines can cause blood clots", "The vaccine is a greater danger to our children's health than COVID itself", "The vaccine will kill you", "The experimental covid vaccine is a death jab", "The covid vaccine causes cancer", "The covid vaccine is harmful for pregnant women and kids", "The vaccine increases health risk", "The vaccine isn't safe", "What are vaccines good for? Nothing, rather it increases risk", "I and many others have medical exemptions", "The vaccine is dangerous for people with medical conditions", "I won't take the vaccine due to medical reasons", "The vaccine has dangerous side effects"
CovidFake	**"Covid-19 disease does not exist"**, "Covid is fake", "covid is a hoax", "covid is a scam", "covid is propaganda", "the pandemic is a lie", "covid isn't real", "I don't think that covid is real", "I don't buy that covid is real", "I don't think there is a pandemic", "I don't think the pandemic is real", "I don't buy that there is a pandemic"
VaxOppression	**"I do not want to be vaccinated because I have freedom of choice"** "Forcing people to take experimental vaccines is oppression", "The vaccine has nothing to do with Covid-19, it's about the vaccine passport and tyranny", "The vaccine mandate is unconstitutional", "I choose not to take the vaccine", "My body my choice", "I'm not against the vaccine but I am against the mandate", "I have freedom to choose not to take the vaccine", "I am free to refuse the vaccine", "It is not about covid, it is about control", "Medical segregation based on vaccine mandates is discrimination", "The vaccine mandate violates my rights", "Falsely labeling the injection as a vaccine is illegal", "Firing over vaccine mandates is oppression", "Vaccine passports are medical tyranny", "I won't let the government tell me what I should do with my body", "I won't have the government tell me what to do"
BigPharmaAnti	**"the vaccine was created only for the profit of pharmaceutical companies"**, "We are the subjects of massive experiments for the Moderna and Pfizer vaccines", "Pharmaceutical companies are corrupt", "The pharmaceutical industry is rotten", "Big Pharma is evil", "How would you trust big pharma with the COVID vaccine? They haven't been liable for vaccine harm in the past", "Covid vaccines are not doing what the pharmaceutical companies promised", "Pharmaceutical companies have a history of irresponsible behavior", "I don't trust Johnson & Johnson after knowing their baby powder caused cancer for decades"
NatImmunityPro	**"natural methods of protection against the disease are better than vaccines"**, "Herd immunity is broad, protective, and durable", "Natural immunity has higher level of protection than the vaccine", "Embrace population immunity", "I trust my immune system", "I have antibodies I do not need the vaccine", "Natural immunity is effective"
VaxAgainstReligion	"The vaccine is against my religion", "The vaccines are the mark of the beast", "The vaccine is a tool of Satan", "The vaccine is haram", "The vaccine is not halal", "I will protect my body from a man made vaccine", "I put it all in God's hands", "God will decide our fate", "The vaccine contains bovine, which conflicts with my religion", "The vaccine contains aborted fetal tissue which is against my religion", "The vaccine contains pork, muslims can't take the vaccine", "Jesus will protect me", "The vaccine doesn't protect you from getting or spreading Covid, God does", "The covid vaccine is another religion"
VaxDoesntWork	**"the vaccine does not work"**, "covid vaccines do not stop the spread", "If the vaccine works, why are deaths so high?", "Why are vaccinated people dying?", "If the vaccine works, why is covid not going away?"
VaxNotTested	**"the vaccine is not properly tested, it has been developed too quickly"**, "Covid-19 vaccines have not been through the same rigorous testing as other vaccines", "The Covid vaccine is experimental", "The covid vaccine was rushed through trials", "The approval of the experimental vaccine was rushed", "How was the vaccine developed so quickly?"
VaxExperimentDogs	"Animal shelters are empty because Dr Fauci allowed experimenting of various Covid vaccines/drugs on dogs and other domestic pets", "Fauci tortures dogs and puppies"
BillGatesMicroChip	"The covid vaccine is a ploy to microchip people", "Bill Gates wants to use vaccines to implant microchips in people", "Globalists support a covert mass chip implantation through the covid vaccine"
VaxFetalTissue	"There is aborted fetal tissue in the Covid Vaccines", "the Covid vaccines contain aborted fetal cells"
VaxMakeYouSterile	"The covid vaccine will make you sterile", "Covid vaccine will affect your fertility"
NoResponsibility	**no one is responsible for the potential side effects of the vaccine**
SwineFluVax	**mentioning the past development of the swine flu vaccine**
VaxResistance	**the vaccine has existed before the Covid-19 epidemic, now there is too much resistance**
ConspiracyTheories	**conspiracy theories, hidden vaccine effects (e.g., chips)**

Table 9: AntiVax arguments and contributed phrases. Arguments that were added during interaction are shown in blue. Arguments that were removed are shown in red. The original definitions/examples are presented in bold.

Arguments	Contributed Phrases
GovTrust	"We trust the government", "The government cares for people", "We are thankful to the government for the vaccine availability", "Hats off to the government for tackling the pandemic", "It is a good thing to be skeptical of the government, but they are right about the covid vaccine", "It is a good thing to be skeptical of the government, but they haven't lied about the covid vaccine", "The government can be corrupt, but they are telling the truth about the covid vaccine", "The government can be corrupt, but they are not lying about the covid vaccine"
VaxSafe	"The vaccine is safe", "Millions have been vaccinated with only mild side effects", "Millions have been safely vaccinated against covid", "The benefits of the vaccine outweigh its risks", "The vaccine has benefits", "The vaccine is safe for women and kids", "The vaccine won't make you sick", "The vaccine isn't dangerous", "The vaccine won't kill you", "The covid vaccine isn't a death jab", "The covid vaccine doesn't harm women and kids"
CovidReal	"Covid is real", "I trust science", "Covid death is real", "The science doesn't lie about covid", "Scientist know what they are doing", "Scientist know what they are saying", "Covid hospitalizations are on the rise", "Covid hospitalizations are climbing as fourth stage surge continues", "Covid's death toll has grown faster", "Covid is not a hoax", "The pandemic is not a lie", "The pandemic is not a lie, hospitalizations are on the rise"
VaxNotOppression	"The vaccine mandate is not oppression because vaccines lower hospitalizations and death rates", "The vaccine mandate is not oppression because it will help to end this pandemic", "The vaccine mandate will help us end the pandemic", "We need a vaccine mandate to end this pandemic", "I support vaccine mandates", "If you don't get the vaccine based on your freedom of choice, don't come crawling to the emergency room when you get COVID", "If you refuse a free FDA-approved vaccine for non-medical reasons, then the government shouldn't continue to give you free COVID tests", "You are free not to take the vaccine, businesses are also free to deny you entry", "You are free not to take the vaccine, businesses are free to protect their customers and employees", "If you choose not to take the vaccine, you have to deal with the consequences", "If it is your body your choice, then insurance companies should stop paying for your hospitalization costs for COVID"
BigPharmaPro	"I trust the science and pharmaceutical research", "Pharmaceutical companies are not hiding anything", "The research behind covid vaccines is public", "The Pfizer vaccine is saving lives", "The Moderna vaccines are helping stop the spread of covid", "The Johnson and Johnson vaccine was created to stop covid", "Pharmaceutical companies are seeking FDA approval", "Pharmaceutical companies are following standard protocols"
NatImmunityAnti	"Only the vaccine will end the pandemic", "Vaccines will allow us to defeat covid without death and sickness", "The vaccine has better long term protection than to natural immunity", "Natural immunity is not effective", "Natural immunity would require a lot of people getting sick", "Experts recommend the vaccine over natural immunity"
VaxReligionOk	"The vaccine is not against religion, get the vaccine", "No religion ask members to refuse the vaccine", "Religious exemptions are bogus", "When turning in your religious exemption forms for the vaccine, remember ignorance is not a religion", "Disregard for others' lives isn't part of your religion", "Jesus is trying to protect us from covid by divinely inspiring scientists to create vaccines"
VaxWorks	"The vaccine works", "Vaccines do work, ask a doctor or consult with an expert", "The covid vaccine helps to stop the spread", "Unvaccinated people are dying at a rapid rate from Covid-19", "There is a lot of research supporting that vaccines work", "The research on the covid vaccine has been going on for a long time"
VaxTested	"Covid vaccine research has been going on for a while", "Plenty of research has been done on the covid vaccine", "The technologies used to develop the Covid-19 vaccines have been in development for years to prepare for outbreaks of infectious viruses", "The testing processes for the vaccines were thorough didn't skip any steps", "The vaccine received FDA approval"
ProVax	**positive attitude**

Table 10: ProVax arguments and contributed phrases. Arguments that were added during interaction are shown in blue. Arguments that were removed are shown in red. The original definition/examples are presented in bold.

Cluster	Experts Rationale	New Named Arguments
K-Means 0	Discusses what the vaccine can and cannot do. Emphasis in reducing COVID-19 symptoms in case of infection (“like a bad cold”). Contains tweets with both stances.	VaxLessensSymptoms
K-Means 1	A lot of mentions to political entities. Politicians get in the way of public safety	GovBadPolicies
K-Means 2	A lot of tweets with mentions and links. Not a lot of textual context. Some examples thanking and praising governmental policies. **Theme added upon inspecting similar tweets**	GovGoodPolicies
K-Means 3	Overarching theme related to vaccine rollout. Mentions to pharmacies that can distribute, distribution in certain states, places with unfulfilled vax appointments. **Too broad to create a theme**	-
K-Means 4	Broadcast of vaccine appointments. Which places you can get vaccine appointments at.	VaxAppointments
K-Means 5	“I got my vaccine” type tweets	GotTheVax
K-Means 6	Mixed cluster, not a clear theme in centroid. Two prominent flavors: the vaccine not working and people complaining about those who are scared of vaccine.	VaxDoesntWork UnjustifiedFearOfVax
K-Means 7	Tweets look the same as K-Means 5	-
K-Means 8	Tweets about development and approval of vaccines	VaxApproval
K-Means 9	Tweets related to common vaccine side-effects	VaxSideEffects

Table 11: **First Iteration**: Patterns Identified in Initial Clusters and Resulting Arguments

Cluster	Experts Rationale	New Named Arguments
K-Means 0	Tweets weighting health benefits/risks, but different arguments. (e.g. it works, doesn’t work, makes things worse...) **Too broad to create a theme**.	-
K-Means 1	Messy cluster, relies on link for information.	-
K-Means 2	Relies on link for information.	-
K-Means 3	A lot of mentions to government lying and misinformation. “misinformation” is used when blaming antivax people. “experts and government are lying” is used on the other side. References to alt-treatments on both sides. **Text lookup “give us the real meds”, “covid meds”**	AntiVaxSpreadMisinfo ProVaxLie AltTreatmentsGood AltTreatmentsBad
K-Means 4	Some examples are a good fit for old theme, VaxDoesntWork. **Other than that no coherent theme.**	-
K-Means 5	Tweets about free will and choice. **Text lookup “big gov”, “free choice”, “my body my choice”** Case “my body my choice” - a lot of mentions to abortion People using covid as a metaphor for other issues.	FreeChoiceVax FreeChoiceOther
K-Means 6	Almost exclusively mentions to stories and news.	-
K-Means 7	Availability of the vaccine, policy. Not judgement of good or bad, but of how well it progresses.	VaxEffortsProgression
K-Means 8	Assign to previous theme GotTheVax	-
K-Means 9	Vaccine side effects. Assign to previous theme, VaxSymptoms	-

Table 12: **Second Iteration**: Patterns Identified in Subsequent Clusters and Resulting Arguments

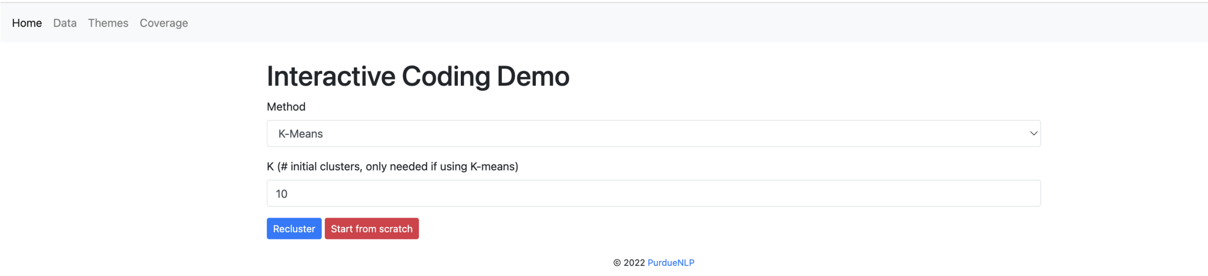


Figure 15: Cluster/Recluster Page

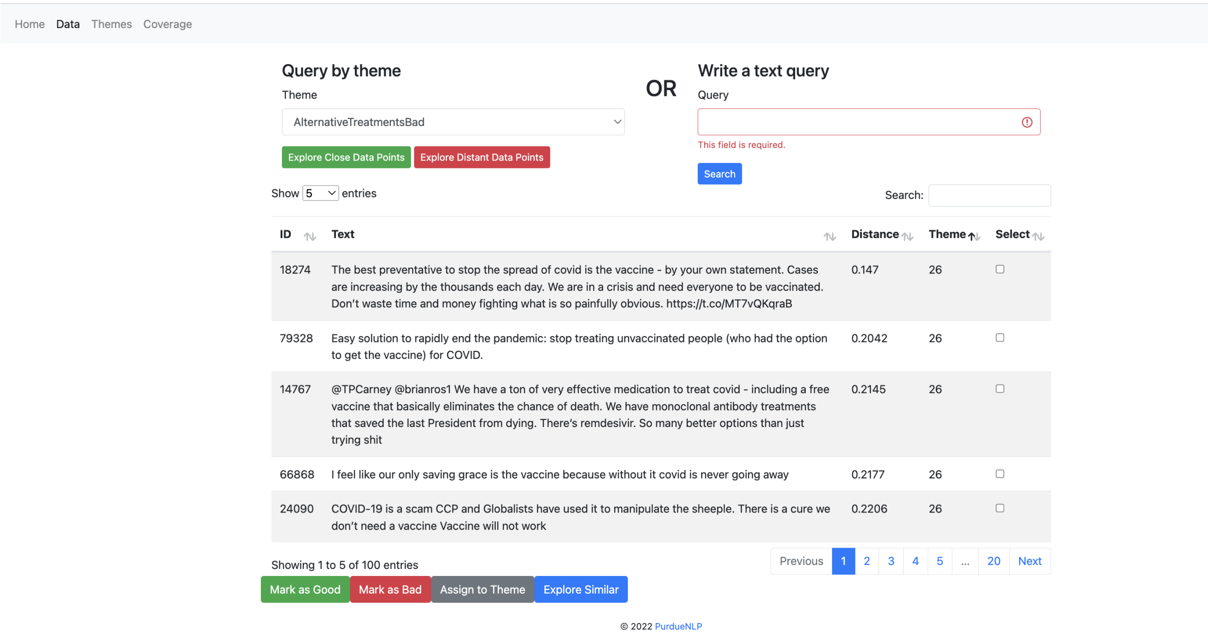


Figure 16: Listing Arguments Page: Named Argument View

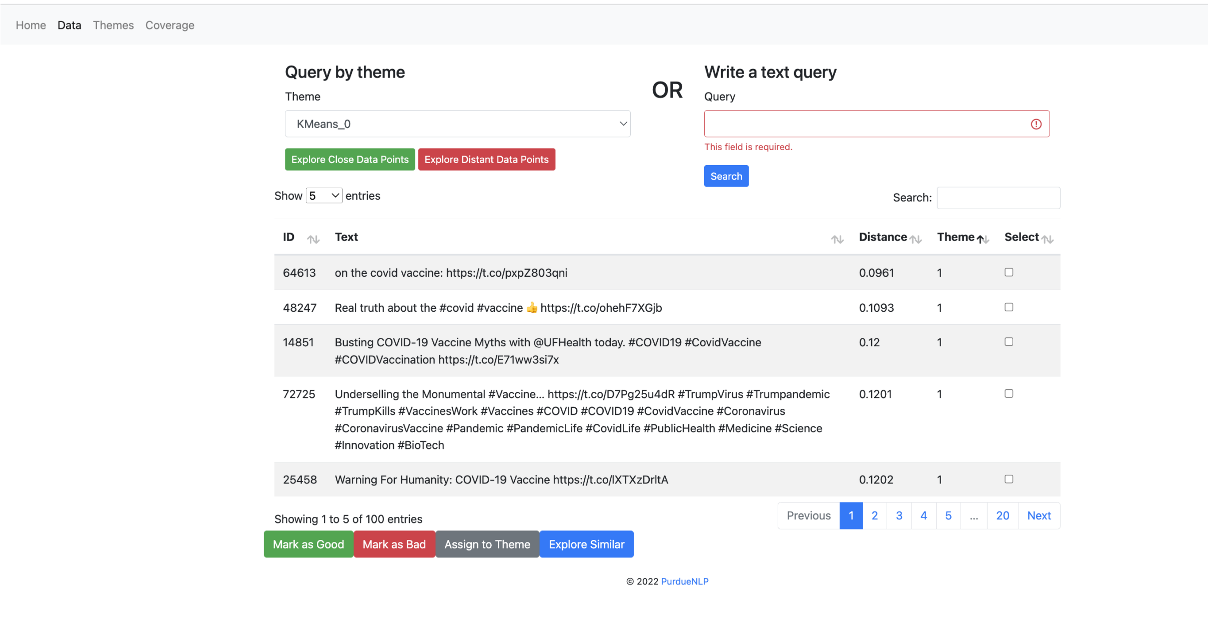


Figure 17: Listing Arguments Page: Unnamed Cluster View

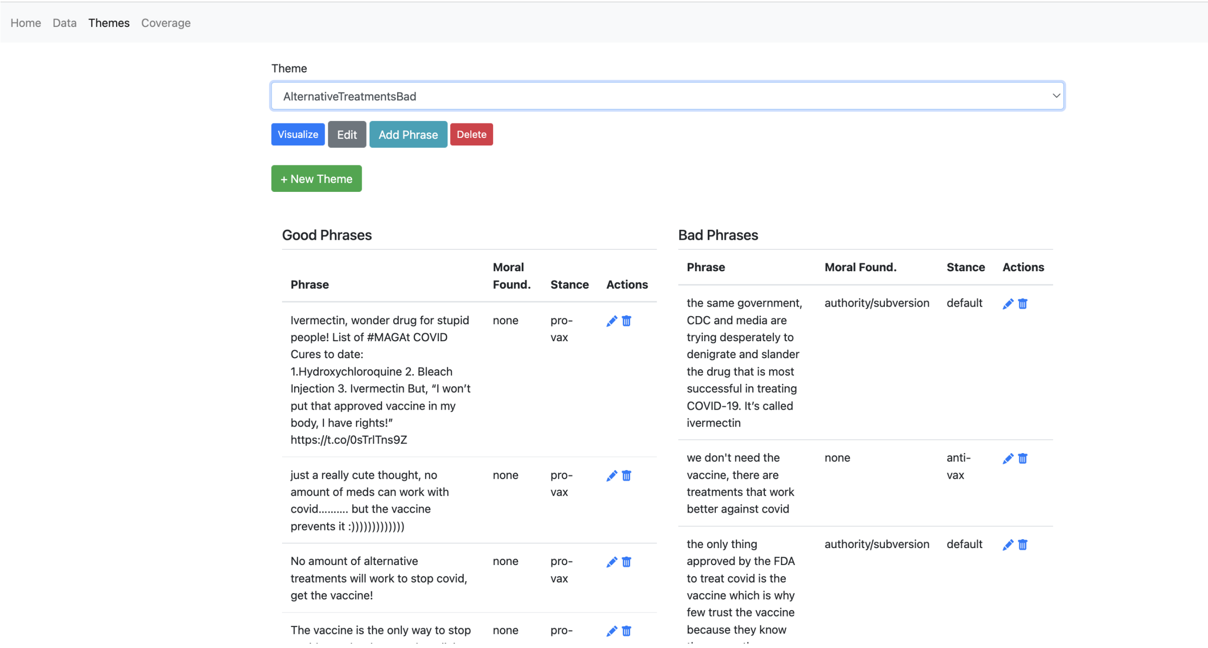


Figure 18: Visualizing Arguments Page

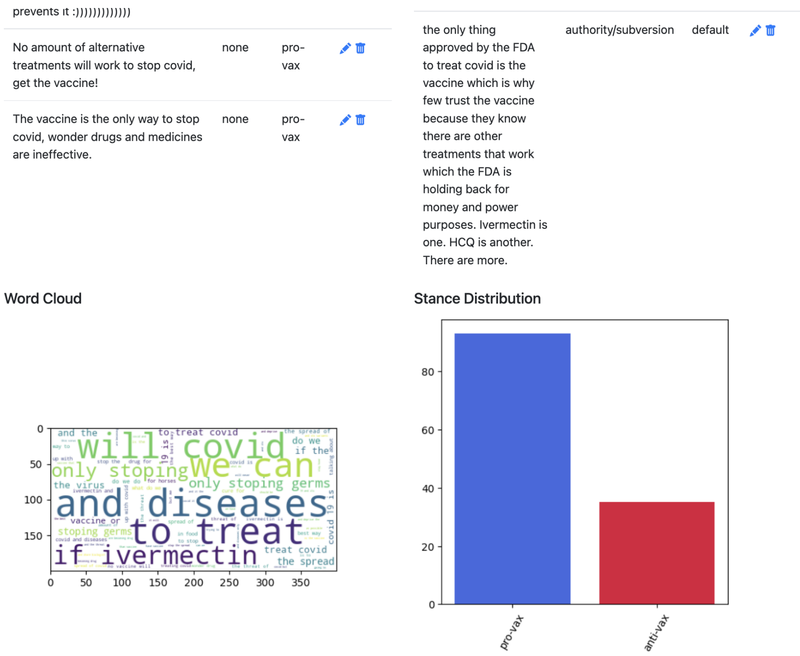
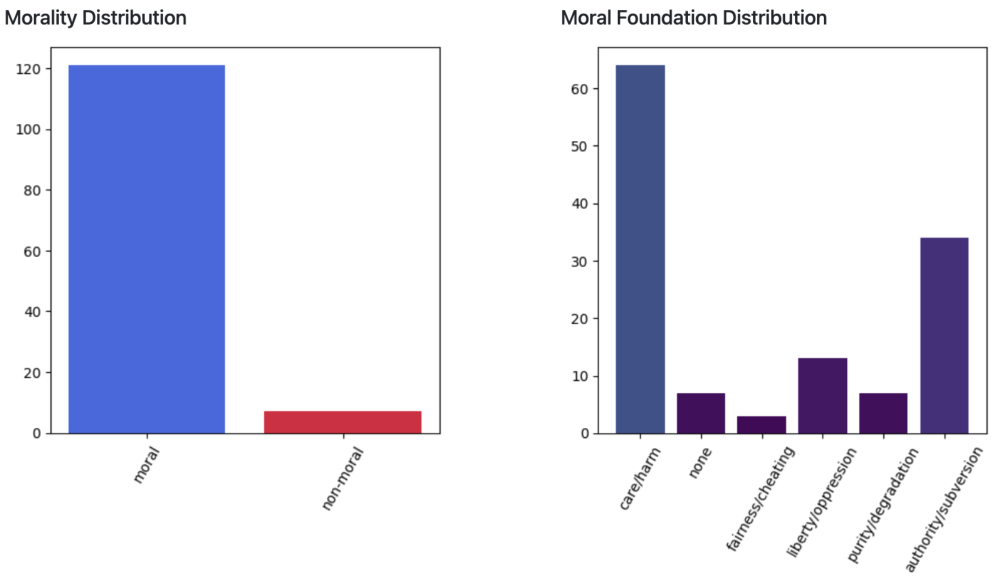


Figure 19: Visualizing Arguments Page: Scroll Down for Local Explanations



Top 10 Positive Entities

entity	count
god	6

Top 10 Negative Entities

entity	count
covid	42

Figure 20: Visualizing Arguments Page: Scroll Down for Local Explanations 2

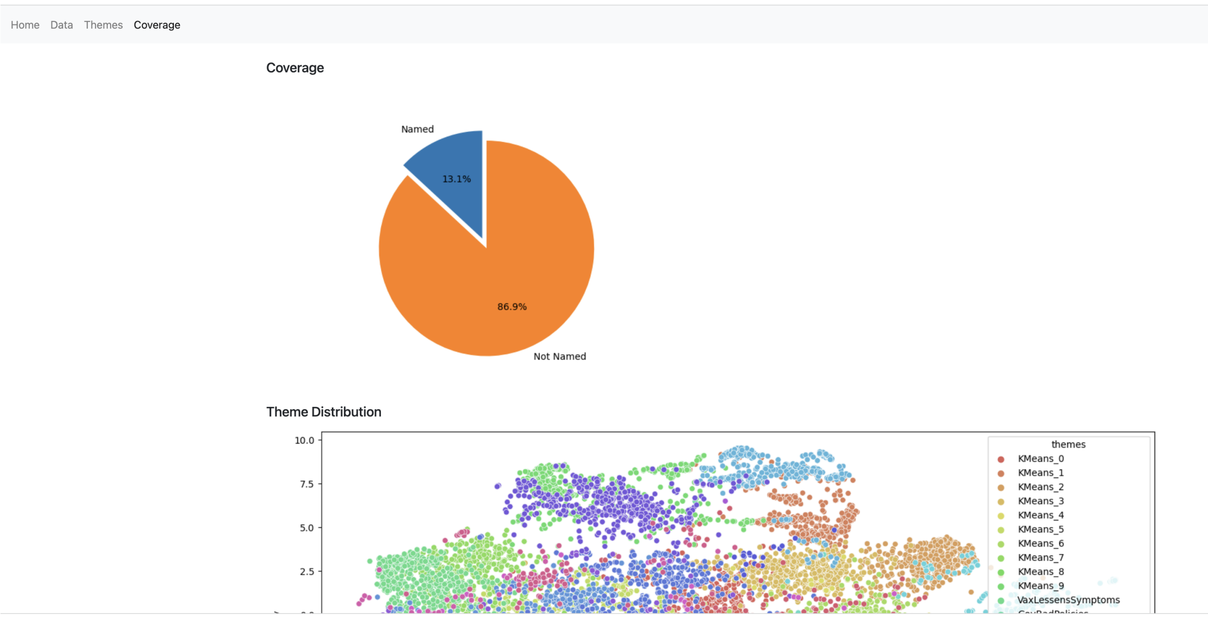


Figure 21: Visualizing Global Explanations Page

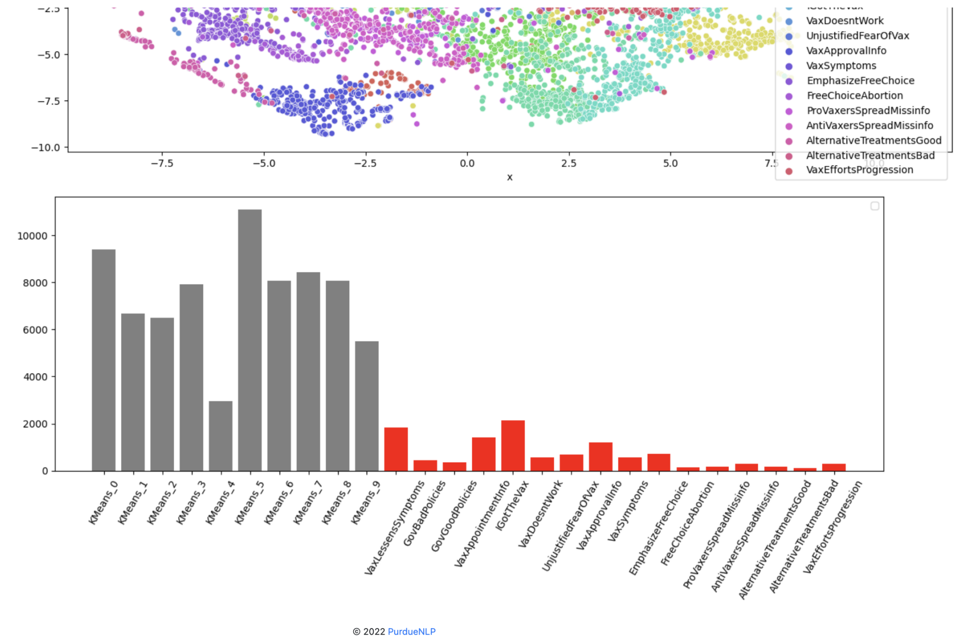


Figure 22: Visualizing Global Explanations Page: Scroll Down for Distributions