A First Step Towards An Interactive Neuro-Symbolic Framework for Identifying Latent Themes in Large Text Collections

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

Abstract

Experts across diverse disciplines are often interested in making sense of large text collections. Traditionally, this challenge is approached either by noisy unsupervised techniques such as topic models, or by following a manual theme discovery process. In this paper, we expand the definition of a theme to account for more than just a word distribution, and include generalized attributes and concepts emerging from the data. Then, we propose an interactive neuro-symbolic framework that receives expert feedback at different levels of abstraction. Our framework strikes a balance between automation and manual coding, allowing experts to maintain control of their study while reducing the manual effort required.

1 Introduction

Researchers and practitioners across diverse academic and professional disciplines are often interested in uncovering latent themes from large text collections. Topic modeling has been the go-to NLP technique to approach this problem (Blei et al., 2003; Boyd-Graber et al., 2017). Despite its wide adoption, this solution is far from perfect, and many efforts have been dedicated to understanding the ways in which topic models can be flawed (Mimno et al., 2011), evaluating their coherence and quality (Stevens et al., 2012; Lau et al., 2014; Röder et al., 2015), and enhancing or replacing them with distributed word representations (Xu et al., 2018; Dieng et al., 2020; Sia et al., 2020). More recently, Hoyle et al. (2021) called the validity of automated topic modeling evaluation techniques into question, by showing that human judgements and automated metrics of quality and coherence do not always agree. Given the noisy landscape surrounding automated topic modeling techniques, manual coding is still prevalent across fields for analyzing nuanced and verbally complex data (Rose, Lennerholt, 2017; Lauer et al., 2018; Antons et al., 2020).

Human-in-the-loop topic modeling approaches aim to address these issues by allowing experts to correct and influence the output of topic models. Given that topics in topic models are defined as distributions over words, these interactive approaches usually receive feedback at the level of individual words (Hu et al., 2011; Lund et al., 2017; Smith et al., 2018). In this paper, we argue that themes emerging from a document collection should not just be defined as a word distribution (similar to a topic model), but generalized attributes and concepts emerging from the data. For example, themes in a dataset about Covid-19 can be characterized by the strength of their relationship to stances about the covid vaccine (e.g. *pro-vax*, *anti-vax*) and moral attitudes towards relevant entities (e.g. *Dr. Fauci* viewed negatively as an entity enabling *cheating*). Working with higher-level abstractions aligns more closely with the way humans approach theme discovery, as it allows them to formulate concepts to generalize from observations to new examples (Rogers, McClelland, 2004), and to deductively draw inferences via conceptual rules and statements (Johnson, 1988). Following the example above, a human could point out that the theme *"The Government is Lying about Covid"* is highly correlated with an *"anti-vax"* stance, and a negative moral sentiment towards *"Dr Fauci"*.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Following this rationale, we suggest an interactive neuro-symbolic approach, aimed to balance unsupervised NLP techniques and manual coding to aid experts in uncovering latent themes from textual repositories. Our main design goal is to provide information to experts, and source feedback from them, at multiple levels of abstraction. Our framework receives a large repository of instances written in natural language, where each instance is associated to a set of observed or predicted attributes. To aid experts in theme discovery, we propose an iterative two-stage machine-in-the-loop framework. In the first stage, we provide the experts with an automated partition of the data and visualizations of the attribute distribution. Then, we have a group of experts work together using a graphical user interface to explore the partitions and identify coherent themes, providing limited feedback both at the text-level and at the attribute-level. In the second stage, the data is re-arranged according to the user feedback. We employ a neuro-symbolic inference process to incorporate the feedback and map instances to the discovered themes. Then, a re-partitioning step is performed on the unassigned instances, and the process is repeated.

As a case study, we focus on Twitter discussions about two polarized topics: the Covid-19 vaccine and immigration. For each topic, we recruit a group of experts and perform two rounds of our two-stage iterative process. Our experiments show that our framework can be used to uncover a set of themes that cover a large portion of the discussion, and that the resulting mapping from tweets to themes is fairly accurate with respect to human judgements.

2 Framework Overview

We propose an iterative two-stage framework that combines interactive interfaces, qualitative methods and neuro-symbolic modeling to assist experts in characterizing large textual collections. We define large textual collections as repositories of textual instances (e.g. tweets, posts, documents) where each instance is associated with a set of annotated or predicted attributes.

In the first stage, our framework automatically proposes an initial partition of the data, such that instances 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 instances, and using open text queries. As the group of experts interact with the data through the interface, they work together following an inductive thematic analysis approach to identify and code the patterns that emerge within the partitions (Braun, Clarke, 2012). Next, they group the identified patterns into general themes, and instantiate them using the interface. Although intuitively we could expect a single cluster to result in a single theme, note that this is not enforced. Experts maintain full freedom as to how many themes they instantiate, if any. Once a theme is created, experts are provided with a set of operations to explain the themes using natural language, select good example instances, write down additional examples, and input or correct supporting attributes. The tool and full set of operations are outlined in Appendix. B.

In the second stage, our framework finds a mapping between the full set of instances and the themes instantiated by the experts. We use the information contributed by the experts in the form of examples and attributes, and learn to map instances to themes. We experiment with two mapping procedures: a nearest neighbors approach that leverages distances in the embedding space between themes and instances, and the proposed ***neuro-symbolic procedure*** that, in addition to the embeddings, considers the additional attributes and judgements provided by the experts. We allow instances to remain unassigned if there is not a good enough match. Following this step, we re-partition all the unassigned instances for a subsequent round of interaction.

Neuro-Symbolic Mapping We used DRaiL (Pacheco, Goldwasser, 2021), a neuro-symbolic modeling framework to design a mapping procedure. Our main goal is to condition new theme assignments not only on the embedding distance between instances and good/bad examples, but also leverage the additional judgements provided by experts using the *"Adding or Correcting Attributes"* procedure. For example, when analyzing the corpus about the Covid-19 vaccine, experts could point out that 80% of the good examples for theme *"Natural Immunity is Effective"* have a clear *anti-vaccine* stance. We could use this information to introduce inductive bias into our mapping procedure, and potentially capture cases where the embedding distance does not provide enough information. DRaiL uses weighted first-order logic rules to express decisions and dependencies between different decisions, which define a probabilistic graphical model. In Fig. 1 we outline

$$\texttt{Inst(i)} \Rightarrow \texttt{Theme(i,t)}$$
$$\texttt{Inst(i)} \Rightarrow \texttt{Attr(i,a)}$$

(a) First-Order Factors

$$\texttt{Inst(i)} \wedge \texttt{Attr(i,a)} \Rightarrow \texttt{Theme(i,t)}$$
$$\texttt{Inst(i)} \wedge \texttt{Attr(i,a}_1\texttt{)} \Rightarrow \texttt{Attr(i,a}_2\texttt{)}$$

(b) Higher-Order Factors

$$\texttt{Inst(i)} \wedge \texttt{Theme(i,t)} \wedge (\texttt{t} \neq \texttt{t}')$$
$$\Rightarrow \neg\texttt{Theme(i,t)}$$

(c) Constraints

Figure 1: DRaiL Rules

the rules introduced. The first set of rules define first order factors, encoding the probability of an instance being mapped to each theme and attribute. We create one template for each theme `t` and attribute `a`, and they correspond to binary decisions (e.g. whether instance `i` mentions theme `t`). Then, we introduce two sets of higher order factors to encode dependencies between each attribute and theme assignment (e.g. probability of theme *"Natural Immunity is Effective"* given that instance has attribute *"anti-vax"*), and between pairs of attributes (e.g. probability of attributes *"anti-vax"* and *"fauci"* co-occurring). Finally, we have a constraint discouraging an instance from having more than one theme assignment.

Our goal is to learn a weight for each rule that captures the probability of that rule being active. Each entity and relation in DRaiL is tied to a neural architecture that is used to learn a distributed representation for it. In this paper, we use a BERT encoder (Devlin et al., 2019) to represent instances, and 1-layer feed-forward networks with ReLU activations over their 1-hot encodings to represent themes and attributes. All relations were encoded as 1-layer feed-forward networks with ReLU activations. Then, parameters for relation and entity encoders, as well as rule weights are learned jointly. The collection of rules represents the global decision, and the solution is obtained by running a maximum a posteriori (MAP) inference procedure. Given that horn clauses can be expressed as linear inequalities corresponding to their disjunctive form, the MAP inference problem can be written as a linear program. DRaiL supports both locally and globally normalized structured prediction objectives. Throughout this paper, we used the locally normalized objective. For details about the learning procedure, we refer the reader to the original paper (Pacheco, Goldwasser, 2021). To generate data for learning the DRaiL model, we take the $K = 100$ closest instances for each good/bad example provided by the experts. Good examples will serve as positive training data. For negative training data, we take the contributed bad examples, as well as good examples for other themes and attributes. Once the weights are learned, we run the inference procedure over the full corpus.

3 Case Studies

We explore two case studies involving discussions on social media: (1) The Covid-19 vaccine discourse in the US, and (2) The immigration discourse in the US, the UK and the EU. For the Covid-19 case, we build on the corpus of 85K tweets released by Pacheco et al. (2022). All tweets in this corpus were posted by users located in the US, are uniformly distributed between Jan. and Oct. 2021, and contain predictions for vaccination stance (e.g. pro-vax, anti-vax) and moral foundations (e.g. fairness/cheating, care/harm, etc.) (Haidt, Graham, 2007). For the immigration case, we build on the corpus of 2.66M tweets released by Mendelsohn et al. (2021). All tweets in this corpus were posted by users located in the US, the UK and the EU, written between 2018 and 2019, and contain predictions for three different frame typologies: narrative frames (e.g. episodic, thematic) (Iyengar, 1991), generic policy frames (e.g. economic, security and defense, etc.) (Card et al., 2015), and immigration-specific frames (e.g. victim of war, victim of discrimination, etc.) (Benson, 2013; Hovden, Mjelde, 2019). Details about the framing typologies can be found in the original publications.

Our main goal in these case studies is to use the framework introduced in Sec. 2 to identify prominent themes in each of these corpora. To do this, we recruited a group of six experts in Natural Language Processing and Computational Social Science, four male and two female, within the ages of 25 and 45. The group of experts included advanced graduate students, postdoctoral researchers and faculty. Our studies are IRB approved, and we follow their protocols. For each corpus, we performed two consecutive sessions with three experts. Each session lasted a total of one hour. In Appendix A, we describe in detail the qualitative thematic analysis process and all of the patterns identified and coded by the experts at each step of the process.

Coverage vs. Mapping Quality: We evaluated the trade-off between coverage (*how many tweets we can account for with the discovered themes*) and mapping quality (*how good we are at mapping tweets*

Iter.	Ground Method	Covid Vaccine # Thm	Covid Vaccine Cover	Covid Vaccine Purity	Covid Vaccine Approx F1	Immigration # Thm	Immigration Cover	Immigration Purity	Immigration Approx F1
Baseline	LDA	9	39.8	63.72	-	13	26.8	57.14	-
1	NNs		9.3	68.81	85.71		11.1	58.44	70.54
	NeSym		**13.7**	**75.69**	**87.50**		**16.4**	**63.89**	**85.29**
Baseline	LDA	16	26.1	65.02	-	19	18.3	57.94	-
2	NeSym		**21.3**	**69.49**			**14.8**	**64.28**	

Table 1: **Dataset Coverage and Average Attribute Purity**. For LDA, we assigned a tweet to its most probable topic if the probability was ≥ 0.5.

to themes). To generate candidates for each theme, we only consider the top 25% of instances in the full dataset that are closest to it in the embedding space. General results are outlined in Fig 1. Given our hypothesis that themes can be characterized by the strength of their relationship to high-level arguments and concepts, we consider mappings to be better if they are more cohesive. In the Covid case, we expect themes to have strong relationships to vaccination stance and moral foundations. In the Immigration case, we expect themes to have strong relationships to the framing typologies. To measure this, we define a theme purity metric for each attribute. For example, for stance this is defined as: $Purity_{stance} = \frac{1}{N} \sum_{t \in Themes} \max_{s \in Stance} |t \cap s|$

In other words, we take each theme cluster and count the number of data points from the most common stance value in said cluster (e.g. the number of data points that are *"anti-vax"*). Then, we take the sum over all theme clusters and divide by the number of data points. We do this for every attribute, and average them to obtain the final averaged attribute purity. We look at the average attribute purity for our mappings at each iteration in the interaction process. In addition to the theme purity, we look at the resulting coverage (e.g. percentage of tweets that were assigned to a theme theme). We can see that the NeSym procedure results in higher purity with respect to the Nearest Neighbors procedure, even when significantly increasing coverage. This is unsurprising, as our method is designed to take advantage of the relationship between themes and attributes. Additionally, we include a topic modeling baseline that does not involve any interaction, and find that interactive themes result in considerably higher purity partitions than topics obtained using LDA, even when LDA covers more instances. Details the LDA implementation used can be found in Appendix B.4.

To approximate F1 for assignment quality, we sub-sampled a set of 200 mapped tweets for each scenario and validated them manually. For the first iteration of Covid-19, we find that the approximated performance of the Neuro-Symbolic mapping is better (+2 points) than the approximated mapping for Nearest Neighbors, while increasing coverage x1.5. For immigration, we have an even more drastic result, having an approximate 15 point increase at a similar coverage gain. In both cases, experts were able to increase the number of themes in subsequent iterations[1]. While the coverage increased in the second iteration for Covid, it decreased slightly for Immigration. For Covid, most of the coverage increase can be attributed to a single, very general theme ("*Vax Efforts Progression*"). In the case of Covid, this large jump in coverage is accompanied by a slight decrease in mapping performance. In the case of Immigration, we have the opposite effect. As the coverage decreases the performance improves, suggesting that the mapping gets stricter. These results confirm the expected trade-off between coverage and performance. Note that we do not perform this manual analysis for LDA, as the topics resulting from LDA are not named, making manual validation more difficult.

4 Summary

We presented a neuro-symbolic framework for uncovering latent themes in text collections. Our framework expands the definitions of a theme to account for attributes and concepts that generalize beyond word co-occurrence patterns, and suggests an interactive protocol that allows human experts to interact with the data and provide feedback at different levels of abstraction. We performed a preliminary evaluation of our framework using two case studies and different groups of experts, and contrasted against the output of traditional topic models. While the experiments in this paper look at short texts, our framework can be easily extended to deal with other types of textual repositories.

[1] Due to effort required and cost, we only do a subsequent interactive session over the NeSym mapping.

References

Antons David, Grünwald Eduard, Cichy Patrick, Salge Oliver. The application of text mining methods in innovation research: current state, evolution patterns, and development priorities // RD Management. 04 2020. 50.

Benson Rodney. Shaping Immigration News: A French-American Comparison. 2013. (Communication, Society and Politics).

Blei David M., Ng Andrew Y., Jordan Michael I. Latent Dirichlet Allocation // J. Mach. Learn. Res. mar 2003. 3, null. 993–1022.

Applications of Topic Models. // . 2017.

Thematic analysis. // . 01 2012. 57–71.

Card Dallas, Boydstun Amber E., Gross Justin H., Resnik Philip, Smith Noah A. The Media Frames Corpus: Annotations of Frames Across Issues // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Beijing, China: Association for Computational Linguistics, VII 2015. 438–444.

Devlin Jacob, Chang Ming-Wei, Lee Kenton, Toutanova Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, VI 2019. 4171–4186.

Dieng Adji B., Ruiz Francisco J. R., Blei David M. Topic Modeling in Embedding Spaces // Transactions of the Association for Computational Linguistics. 2020. 8. 439–453.

Haidt Jonathan, Graham Jesse. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize // Social Justice Research. 2007. 20, 1. 98–116.

Hovden Jan Fredrik, Mjelde Hilmar. Increasingly Controversial, Cultural, and Political: The Immigration Debate in Scandinavian Newspapers 1970–2016 // Javnost - The Public. 2019. 26, 2. 138–157.

Hoyle Alexander, Goel Pranav, Hian-Cheong Andrew, Peskov Denis, Boyd-Graber Jordan, Resnik Philip. Is Automated Topic Model Evaluation Broken? The Incoherence of Coherence // Advances in Neural Information Processing Systems. 34. 2021. 2018–2033.

Hu Yuening, Boyd-Graber Jordan, Satinoff Brianna. Interactive Topic Modeling // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA: Association for Computational Linguistics, VI 2011. 248–257.

Iyengar Shanto. Is anyone responsible? : how television frames political issues. Chicago: University of Chicago Press, 1991. (American politics and political economy series).

Jin Xin, Han Jiawei. K-Means Clustering // Encyclopedia of Machine Learning. Boston, MA: Springer US, 2010. 563–564.

Johnson Ralph H. Gilbert Harman Change in View: Principles of Reasoning (Cambridge, MA: MIT Press 1986). Pp. ix 147. // Canadian Journal of Philosophy. 1988. 18, 1. 163–178.

Lau Jey Han, Newman David, Baldwin Timothy. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality // Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden: Association for Computational Linguistics, IV 2014. 530–539.

Lauer Claire, Brumberger Eva, Beveridge Aaron. Hand Collecting and Coding Versus Data-Driven Methods in Technical and Professional Communication Research // IEEE Transactions on Professional Communication. 2018. 61, 4. 389–408.

Lund Jeffrey, Cook Connor, Seppi Kevin, Boyd-Graber Jordan. Tandem Anchoring: a Multiword Anchor Approach for Interactive Topic Modeling // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics, VII 2017. 896–905.

Maaten Laurens van der, Hinton Geoffrey. Visualizing Data using t-SNE // Journal of Machine Learning Research. 2008. 9. 2579–2605.

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

Mendelsohn Julia, Budak Ceren, Jurgens David. Modeling Framing in Immigration Discourse on Social Media // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, VI 2021. 2219–2263.

Mimno David, Wallach Hanna, Talley Edmund, Leenders Miriam, McCallum Andrew. Optimizing Semantic Coherence in Topic Models // Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, UK.: Association for Computational Linguistics, VII 2011. 262–272.

Pacheco Maria Leonor, Goldwasser Dan. Modeling Content and Context with Deep Relational Learning // Transactions of the Association for Computational Linguistics. 2021. 9. 100–119.

Pacheco Maria Leonor, Islam Tunazzina, Mahajan Monal, Shor Andrey, Yin Ming, Ungar Lyle, Goldwasser Dan. A Holistic Framework for Analyzing the COVID-19 Vaccine Debate // Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, United States: Association for Computational Linguistics, VII 2022. 5821–5839.

Rehurek Radim, Sojka Petr. Gensim–python framework for vector space modelling // NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic. 2011. 3, 2.

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

Röder Michael, Both Andreas, Hinneburg Alexander. Exploring the Space of Topic Coherence Measures // Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. New York, NY, USA: Association for Computing Machinery, 2015. 399–408. (WSDM '15).

Semantic Cognition: A Parallel Distributed Processing Approach. // . 2004.

Rose Jeremy, Lennerholt Christian. Low cost text mining as a strategy for qualitative researchers // Electronic Journal on Business Research Methods. 04 2017. forthcoming.

Sia Suzanna, Dalmia Ayush, Mielke Sabrina J. Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too! // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020. 1728–1736.

Smith Alison, Kumar Varun, Boyd-Graber Jordan, Seppi Kevin, Findlater Leah. Closing the Loop: User-Centered Design and Evaluation of a Human-in-the-Loop Topic Modeling System // 23rd International Conference on Intelligent User Interfaces. New York, NY, USA: Association for Computing Machinery, 2018. 293–304. (IUI '18).

Stevens Keith, Kegelmeyer Philip, Andrzejewski David, Buttler David. Exploring Topic Coherence over Many Models and Many Topics // Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, Korea: Association for Computational Linguistics, VII 2012. 952–961.

Xu Hongteng, Wang Wenlin, Liu Wei, Carin Lawrence. Distilled Wasserstein Learning for Word Embedding and Topic Modeling // Advances in Neural Information Processing Systems. 31. 2018.

A Thematic Analysis Results

For the Covid-19 study, the clusters for the first iteration of interaction, as well as the coded argumentative patterns and the resulting themes can be observed in Tab. 2. The same content for the second iteration of interaction can be observed in Tab. 3. Tables 4 and 5 outline the patterns discovered by the experts for immigration.

Cluster	Experts Rationale	New Named Themes
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 2: **First Iteration for Covid-19**: Patterns Identified in Initial Clusters and Resulting Themes

Cluster	Experts Rationale	New Named Themes
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 3: **Second Iteration for Covid-19**: Patterns in Subsequent Clusters and Resulting Themes

Table 4 and 5 outline the patterns discovered by the experts for immigration.

Cluster	Experts Rationale	New Named Themes
K-Means 0	Headlines, coverage. Some have an agenda (pro) Others are very academic and research-oriented Opinion pieces.	AcademicDiscussions
K-Means 1	Talking about apprehending immigrants at the border Some report about the border but no stance. Deportation. Leaning negative towards immigrants.	JustifiedDetainmentEnforce
K-Means 2	Less US-centric, more general. Talking about immigration as a global issue Humanitarian issues, mentions to refugees, forced migration Situation in country of origin that motivates immigration Mentions to how the west is responsible The role of the target countries in destabilizing countries Mentions to economic migrants. **Look up for "economic work migrants", "asylum seekers"**	EconomicMigrantsNotAsylumSeekers SituationCountryOfOrigin RoleOfWesternCountries
K-Means 3	About Trump. Trump immigration policy. Politicizing immigration.	TrumpImmiPolicy
K-Means 4	Attacking democrats. A lot of mentions to democrats wanting votes Common threads is democrats are bad	DemocratImmiPolicyBad
K-Means 5	Lacks context, lots of usernames. Not a cohesive theme. Both pro and con, and vague. Some mentions to invasion. **Look for "illegal immigrants invade"** Mentions to caravan, massive exodus of people. Mentions to crime. **Look for immigrants murder, immigrants dangerous.** A lot of tweets linking immigrants to crime	ImmigrantInvasion ImmigrantCrime
K-Means 6	Looks very varied. Not cohesive.	-
K-Means 7	Very cohesive. Mentions to detaining children, families.	DetainingChildren
K-Means 8	All tweets are about the UK and Britain. Both pro and anti immigration. Only common theme is the UK. Almost exclusively policy/politics	UKProImmiPolicy UKAntiImmiPolicy
K-Means 9	Economic cost of immigration. Immigration is bad for the US economy Some about crime, and democrats. Assign to existing themes.	FinacialCostOfImmigration

Table 4: **First Iteration Immigration**: Patterns Identified in Initial Clusters and Resulting Themes

Cluster	Experts Rationale	New Named Themes
K-Means 0	Legal decisions and rulings. Both pro and anti immigration rulings Not a single event, but cohesively talking about rulings	CourtRulings
K-Means 1	The same tweet reworded and tweeted at different people Talks about worker exploitation, and Cesar Chavez. **Look up for "exploitation"**. Mentions to workers and wages **Look up for "cheap labor"**	ImmigrantWorkerExploitation
K-Means 2	Blaming Trump for being irresponsible Criticizing his rhetoric. Mentions to hateful speech About the rhetoric rather than policy. Mentions to racist language Others about policy, added to previous TrumpImmiPolicy theme	CriticizeAntiImmigrantRhetoric
K-Means 3	Nation of immigrants. Identity, we are all immigrants	CountryOfImmigrants
K-Means 4	Organizing. Call to action. Skews pro. language of rights and liberties. We are here, we demand, sign here. **Look up "ACLU", "rights for immigrants"**	ProImmiActivism
K-Means 5	A lot of mentions to numbers and stats. Short URLs. Headlines.	-
K-Means 6	A lot of usernames. Bad policies, criticizing policies on both sides. Send them to either DemocratImmiPolicyBad or TrumpImmiPolicy	-
K-Means 7	Very messy. Links.	-
K-Means 8	European headlines and news. Some about the UK. Send the ones that are relevant to UK policy themes	
K-Means 9	Detention, detention centers, solitary confinement as cruel.	DetainmentCruel

Table 5: **First Iteration Immigration**: Patterns Identified in Initial Clusters and Resulting Themes

B Interactive Tool

To support our interactive framework, we developed a tool for human experts to interact with the textual repositories. The tool is a simple GUI equipped with a finite set of exploratory and intervention operations. *Exploratory operations* allow experts to discover clusters of instances and further explore and partition the space, as well as to evaluate the quality of the discovered clusters and theme-instance mappings. *Intervention operations* allow experts to name the discovered patterns, as well as to provide examples and judgements to improve the quality of the mappings (See Tab. 6). We represent instances using their Sentence BERT embedding (Reimers, Gurevych, 2019). We represent themes using a handful of explanatory phrases and a small set of examples, and calculate their SBERT embeddings. Screenshots showcasing the GUI are also included below.

Operations	Description
Finding Clusters	Experts can find clusters in the space of unassigned instances. To do this, we run a clustering algorithm using the representations described in Sec. **??**. We support the K-means Jin, 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 instances that are close to the query in the embedding space.
Finding Similar Instances	Experts have the ability to select each instance and find other examples that are close in the embedding space.
Listing Themes and Instances	Experts can browse the current list of themes and their mapped instances. Instances are ranked in order of "goodness", corresponding to the similarity in the embedding space to the theme representation. They can be 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 themes. 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 theme distribution, coverage statistics, and t-sne plots Maaten van der, Hinton (2008).

(a) Exploratory Operations

Operations	Description
Adding, Editing and Removing Themes	Experts can create, edit, and remove themes. The only requirement for creating a new theme is to give it a unique name. Similarly, themes can be edited or removed at any point. If any instances are assigned to a theme being removed, they will be moved to the space of unassigned instances.
Adding and Removing Examples	Experts can assign "good" and "bad" examples to existing themes. Good examples are instances that characterize the named theme. 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 mapped instances as "good" or "bad", or they can directly contribute example phrases.
Adding or Correcting Attributes	We allow users to upload additional observed or predicted attributes for each textual instance. For instances and phrases added as "good" and "bad" examples, we allow users to add or edit the values of these attributes. The intuition behind this operation is to collect additional information for learning to map instances to themes.
Mapping Instances to Themes	Experts can toggle the assignment of instances to existing themes. Currently, we support two mapping approaches: a nearest neighbors approach, which relies only on embedding distances, and a neuro-symbolic approach, which makes use of all the provided judgments and features.

(b) Intervention Operations

Table 6: Interactive Operations

B.1 Discovery Operations

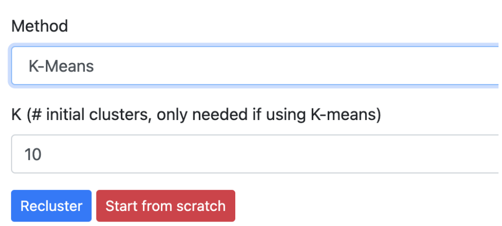


Figure 2: Cluster Instances

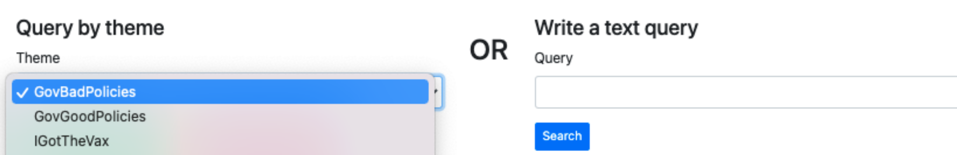
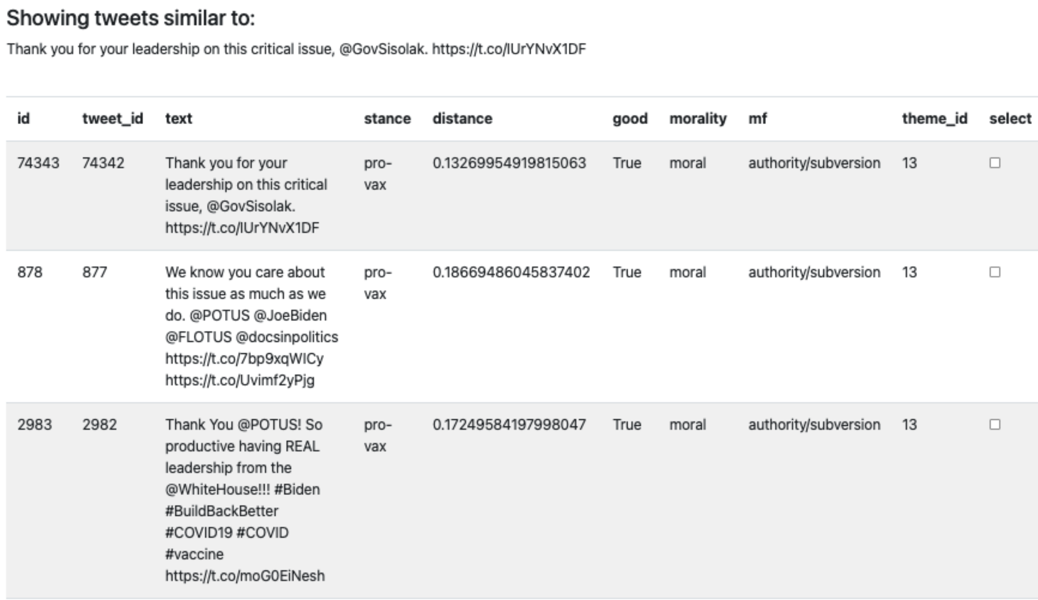


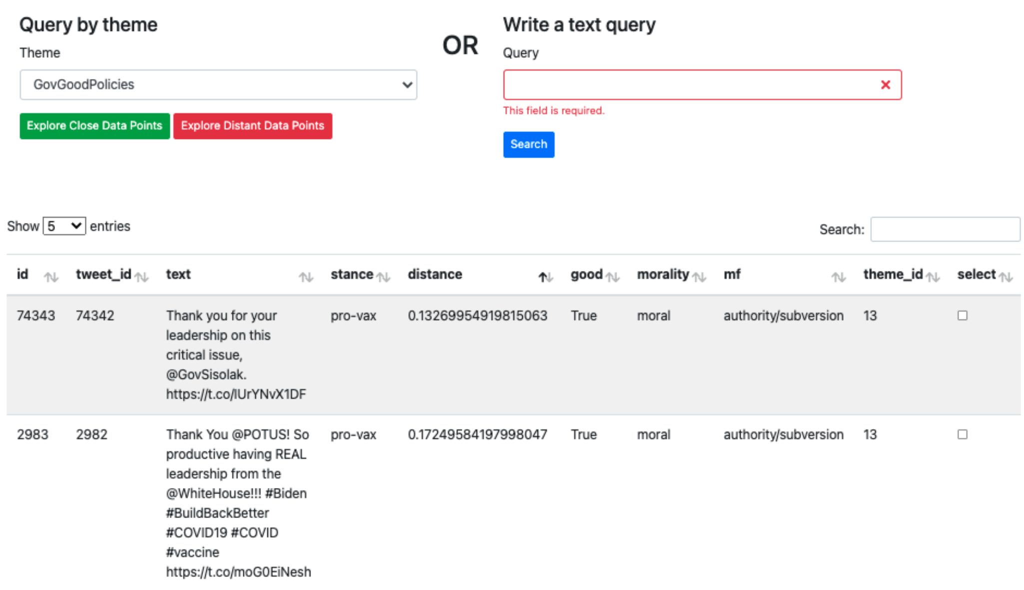
Figure 3: Text-based Queries


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

B.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

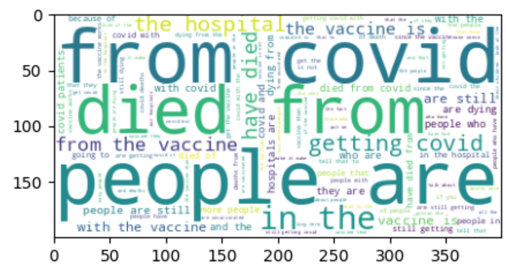


Figure 6: Visualizing Local Explanations: Word Cloud Example for *The Vaccine Doesn't Work*

Top 10 Positive Entities

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

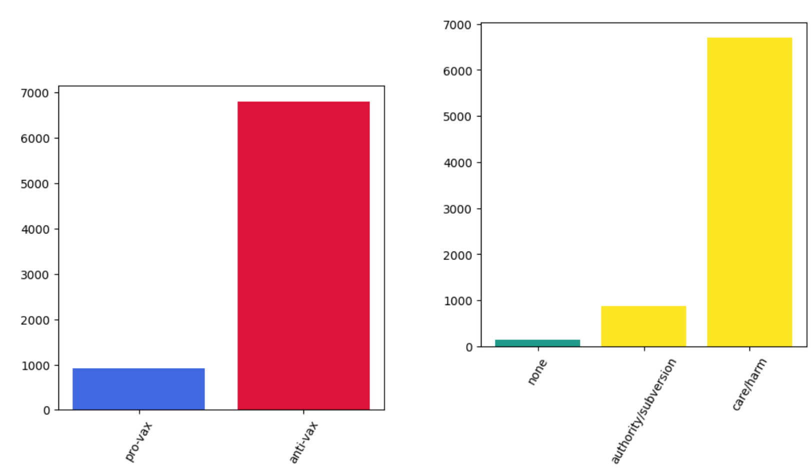
(a) Top Positive Entities

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

(b) Moral Foundation

Figure 8: Visualizing Local Explanations: Attribute Distribution for *The Vaccine Doesn't Work*

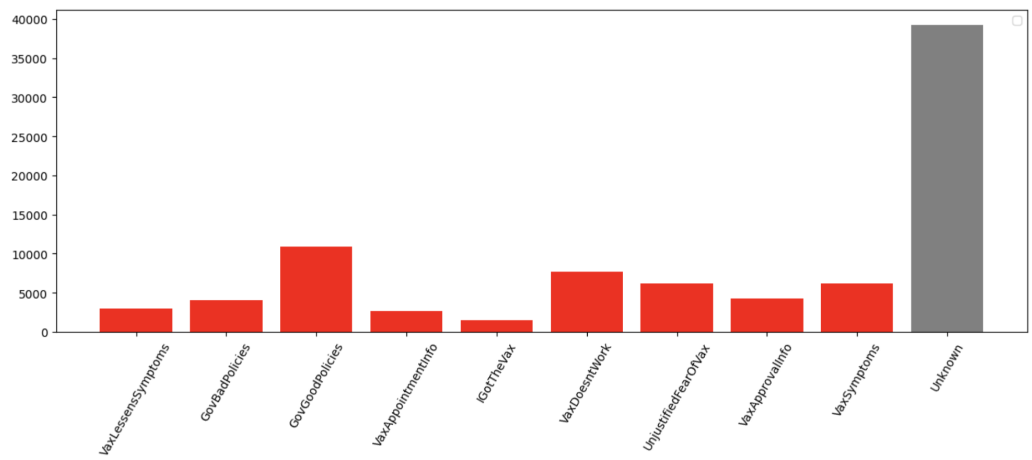


Figure 9: Visualizing Global Explanations: Theme Distribution

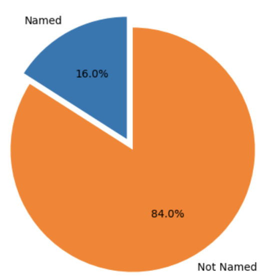


Figure 10: Visualizing Global Explanations: Coverage

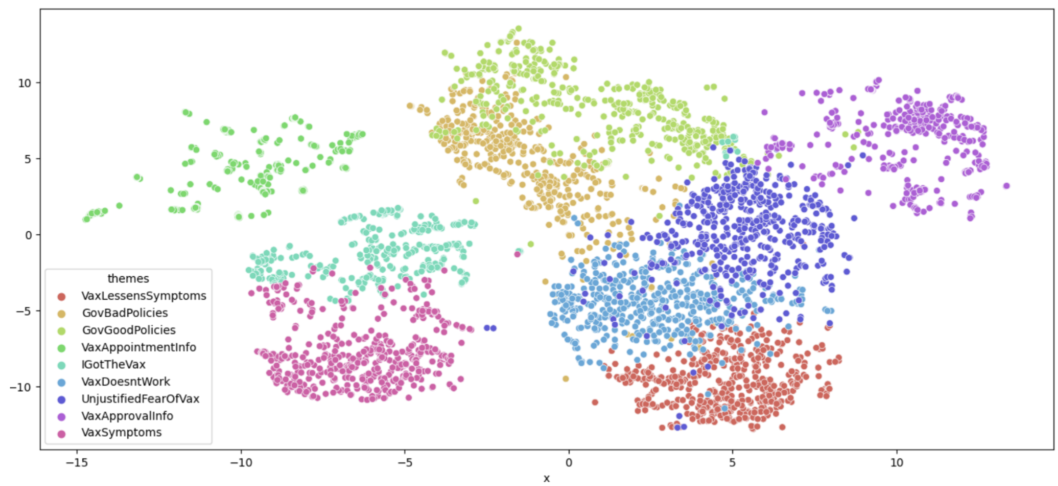


Figure 11: Visualizing Global Explanations: 2D t-SNE

B.3 Intervention Operations

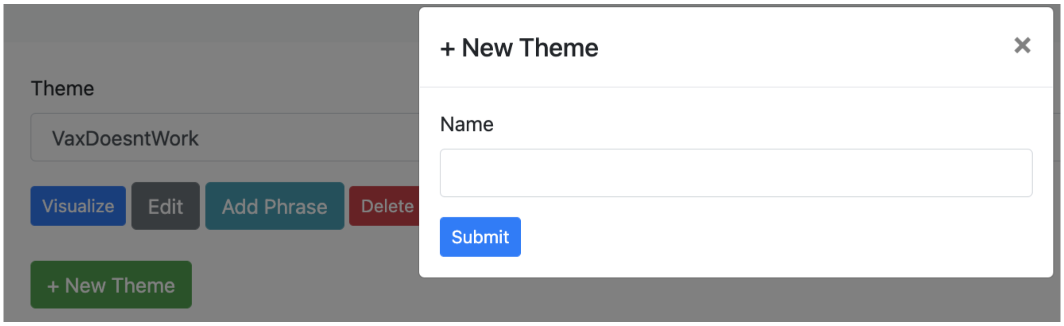


Figure 12: Adding New Themes

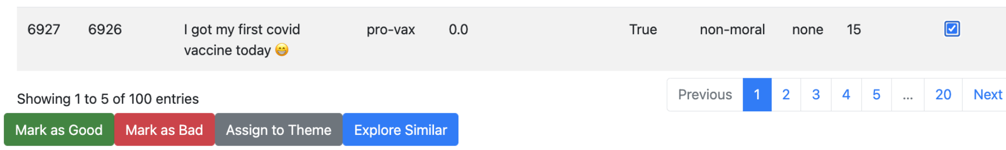


Figure 13: Marking Instances as *Good*

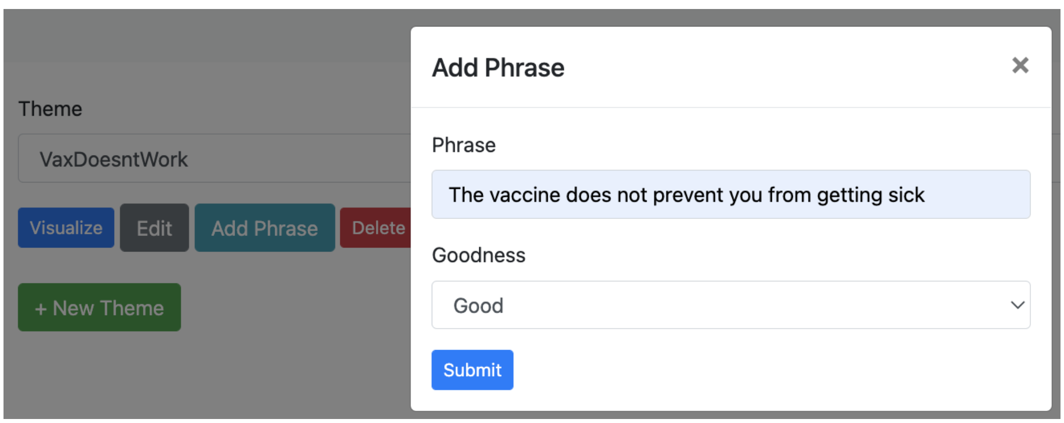


Figure 14: Adding *Good* Examples

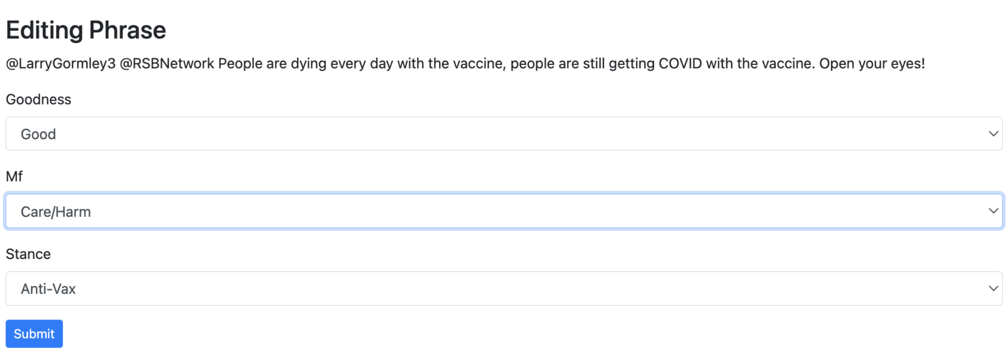


Figure 15: Correcting Attributes - Stances and Moral Foundations

B.4 Topic Modeling Details

To obtain LDA topics, we use the Gensim implementation Rehurek, Sojka (2011) and follow all the prepossessing steps suggested by Hoyle et al. (2021), with the addition of the words covid, vaccin* and immigra* to the list of stopwords.