An Interactive Neuro-Symbolic Framework for Uncovering Latent Themes in Large Text Collections

Anonymous ACL submission

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 and 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 and 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"*.

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 The Framework

We propose an iterative two-stage framework that combines ML/NLP techniques, interactive interfaces and qualitative methods 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 potentially 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 and 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 operations are outlined in Sec. 2.1.

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 allow instances to remain unassigned if there is not a good enough match. We experiment with two mapping procedures: a simple nearest neighbors approach that leverages distances in the embedding space between themes and instances, and a neuro-symbolic procedure that, in addition to the embeddings, considers the additional attributes and judgements provided by the experts. The two procedures are outlined in Sec. 2.3.

2.1 Interactive Tool

To support our interaction protocol, we developed a tool for experts to interact with the language resource. This tool is a simple GUI equipped with a finite set of exploratory and intervention operations. *Exploratory operations* allow uses to discover clusters and further explore and partition the space, and to evaluate the quality of the discovered clusters and the mapped instances. *Intervention operations* allow users to name the discovered patterns, as well as to provide examples and judgements to improve the quality of the partitions. Operations are listed in Tabs. 1 and 2, and demonstrated in App. A.1.

Representing Themes and Instances: We represent example instances using their Sentence-BERT (SBERT) embedding (Reimers and Gurevych, 2019). We represent themes using a handful of explanatory phrases and a small set of examples, and calculate their SBERT embeddings. To measure the closeness between an instance and a theme, we compute the cosine similarity between the instance and all of the explanatory phrases and examples for the theme, and take the maximum similarity score among them. Note that our tool and the operations presented are agnostic of the representation used. The underlying embedding objective, as well as the "closeness" scoring function can be easily replaced.

2.2 Interaction Protocol

We follow a simple protocol where three human coders work together using the operations de-

Operations	Description
Finding Clusters	Experts can find clusters in the space of unassigned instances. 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 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 attribute 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 (van der Maaten and Hinton, 2008).

Table 1: 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 attributes. Both of these approaches are outlined in Sec. 2.3.

Table 2: Intervention Operations

scribed in Sec. 2.1 to discover themes in large textual corpora. In addition to the three coders, each interactive session is guided by one of the authors of the paper, who makes sure the coders are adhering to the process outlined here.

To initialize the system, the coders will start by using the clustering operation to find 10 initial clusters of roughly the same size. During the first session, the coders will inspect the clusters one by one by looking at the examples closest to the centroid. This will be followed by a discussion phase, in which the coders follow an inductive thematic analysis approach to identify repeating patterns and write them down. If one or more cohesive patterns are identified, the experts will create a new theme, name it, and mark a set of good example instances that help in characterizing the named theme. When a pattern is not obvious, coders will explore similar instances to the different statements found. Whenever the similarity search results in a new pattern, the coders will create a new theme, name it, and mark a set of good example instances that helped in characterizing the named theme.

Next, the coders will look at the local theme explanations and have the option to enhance each theme with additional phrases. Note that each theme already contains a small set of representative instances, which are marked as "good" in the previous step. In addition to contributing "good" example phrases, coders will have the option to contribute some "bad" example phrases to push the representation of the theme away from statements that have high lexical overlap with the good examples, but different meaning. Finally, coders will examine each exemplary instance and phrase for the set of symbolic attributes (e.g. stance, sentiment.). In cases where the judgement is perceived as wrong, the coders will be allowed to correct it. In this paper, we assume that the textual corpora include a set of relevant attributes for each instance. In future work, we would like to explore the option of letting coders define attributes on the fly.

2.3 Mapping and Re-clustering

Each interactive session will be followed by a mapping and re-clustering stage. First, we perform the mapping step, in which we assign instances to the themes discovered during interaction. We do not assume that experts have discovered the full space of latent themes. For this reason, we do not try to assign a theme to each and every instance. We expect that the set of themes introduced by the human experts at each round of interaction will cover a subset of the total instances available. Following this step, we will re-cluster all the unassigned instances for a subsequent round of interaction. We evaluate two methods to assign instances to themes:

Nearest Neighbors Mapping Approach: In this approach, each instance is assigned to its closest theme *if and only if* the distance to the closest theme is less than or equal to the distance to its previous cluster *and* the distance to the closest theme is less than or equal to the distance to the theme's bad examples and phrases.

Neuro-Symbolic Mapping Approach: We used

DRaiL (Pacheco and 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 first-order logic rules to express decisions and dependencies between different decisions. We introduce the following rules:

$$\begin{aligned}
t_0 - t_n &: \texttt{Inst(i)} \Rightarrow \texttt{Theme(i,t)} \\
a_0 - a_m &: \texttt{Inst(i)} \Rightarrow \texttt{Attr(i,a)} \\
c_0 - c_{n*m} &: \texttt{Inst(i)} \wedge \texttt{Attr(i,a)} \Rightarrow \texttt{Theme(i,t)} \\
c'_0 - c'_{n*n} &: \texttt{Inst(i)} \wedge \texttt{Theme(i,t)} \wedge (\texttt{t} \neq \texttt{t}') \\
&\quad \Rightarrow \neg\texttt{Theme(i,t)}
\end{aligned}$$

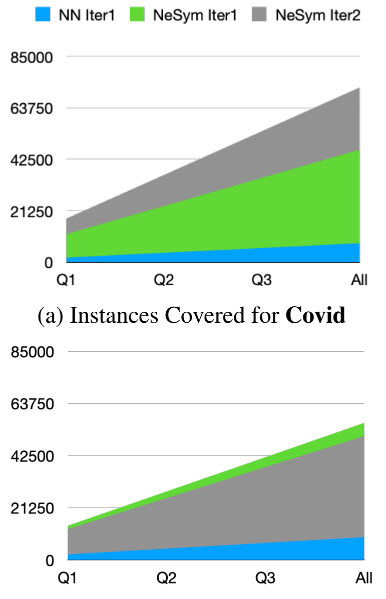
The first set of rules $t_0 - t_n$ and $a_0 - a_m$ map instances to themes and attributes respectively. We create one template for each theme `t` and attribute `a`, and they correspond to binary decisions (e.g. whether instance `i` mentions theme `t`). Then, we introduce two sets of soft constraints: $c_0 - c_{n*m}$ encode the dependencies between each attribute and theme assignment (e.g. likelihood of theme *"Natural Immunity is Effective"* given that instance has attribute *"anti-vax"*). Then, $c'_0 - c'_{n*n}$ discourages an instance from having more than one theme assignment. For each rule, we will learn a weight that captures the strength of that rule being active. Then, a combinatorial inference procedure is run to find the most likely global assignment. Each entity and relation in DRaiL is tied to a neural architecture that is used to learn its weights. In this paper, we use a BERT encoder (Devlin et al., 2019) for all rules. 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. More information about DRaiL can be found in the original paper.

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 and 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 and Mjelde, 2019). Additional details about the datasets and framing typologies can be found the original publications.

Our main goal in these studies is to use the framework introduced in Sec. 2 to identify prominent themes in the corpora introduced above. 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 following the protocol outlined in Sec. 2.2. To evaluate consistency, we did an additional two sessions with a different group of experts for the Covid-19 dataset. Each session lasted a total of one hour. In Appendices A.2, A.3 and A.4, we include large tables enumerating the resulting themes, and describing in detail 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 can we account for with the discovered themes) and mapping quality (how good are we at mapping tweets to themes). Results are outlined in Fig. 1. To do this evaluation, we sub-sampled a



(a) Instances Covered for **Covid**

(c) Instances Covered for **Immigration**

Iter.	Ground. Method	$\leq Q_1$	$\leq Q_2$	$\leq Q_3$	All
1	NNs	89.80	87.50	87.50	85.71
	NeSym	87.50	81.32	75.38	70.66
2	NeSym	85.71	76.92	73.13	68.49

(b) **Covid** Theme F1

Iter.	Ground. Method	$\leq Q_1$	$\leq Q_2$	$\leq Q_3$	All
1	NNs	86.96	76.19	74.19	70.54
	NeSym	85.29	79.07	73.51	70.54
2	NeSym	91.43	83.08	79.15	76.76

(d) **Immigration** Theme

Figure 1: Theme Assignments Where Distance to Theme Centroid $\leq$ Quartile

set of 200 mapped tweets for each scenario, uniformly distributed across themes and similarity to the theme embedding, and validated them manually. The logic behinds this is that we expect mapping performance to degrade the more semantically different the tweets are to the "good" examples and phrases provided by the experts. To achieve this, we look at evaluation metrics at different thresholds using the quartiles with respect to the similarity distribution. Results for the first quartile (Q_1) correspond to the 25% most similar instances. For the second quartile (Q_2), to the 50% most similar instances, and for the third quartile (Q_3), to the 75% most similar instances.

For the first iteration of Covid-19, we find that the approximated performance of the Neuro-Symbolic mapping at Q1 is better (+2 points) than the approximated full 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 theme ("*Vax Efforts Progression*"), which accounts for 20% of the mapped data. 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. Depending on the needs of the final applications, experts could adjust their confidence thresholds. To perform a fine-grained error analysis, we looked at the errors made by the model during manual validation. In Fig. **??** we show the confusion matrix for the Covid case. We find that the performance varies a lot, with some themes being more accurate than others. In some cases, we are good at capturing the general meaning of the theme but fail at grasping the stance similarities (e.g. *Anti Vax Spread Missinfo* gets confused with *Pro Vax Lie*, where the difference is on who is doing the lying). In other cases, we find that themes that are close in meaning have some overlap (e.g. *Alt Treatments* with *Vax Doesn't Work*). We also find that unambiguous, neutral themes like *Vax Appointments*, *Got The Vax* and *Vax Efforts Progression* have the highest performance.

Lastly, we observe that for some errors, none of the existing themes are appropriate (Last row: *Other*), suggesting that there are still themes that have not been discovered. Upon closest inspection, we found that the majority of these tweets are among the most distant from the theme embedding. The full distribution of *"Other"* per interval can be ob-

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

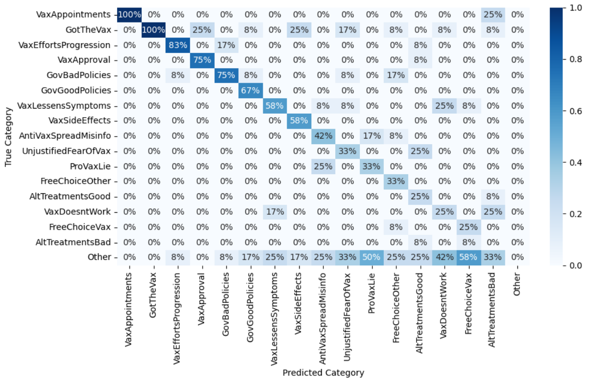


Figure 2: **Confusion matrix for Covid after second iteration**. Values are normalized over the predicted themes (cols), and sorted best to worst.

served in App. A.6. We also include the confusion matrix for immigration in App. A.6.

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. In Tab. 3 we look at the average attribute purity for our mappings at each iteration in the interaction process. 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 higher purity partitions than topics obtained using LDA. Details about the steps taken to obtain LDA topics can be found in App. A.5.

Iter.	Ground	Covid Vaccine			Immigration		
	Method	# Thm	Cover	Purity	# Thm	Cover	Purity
Baseline	LDA		39.8	63.72		26.8	57.14
1	NNs	9	9.3	68.81	13	11.1	58.44
	NeSym		**54.3**	**69.97**		**65.8**	**61.72**
Baseline	LDA	16	26.1	65.02	19	18.3	57.94
2	NeSym		**84.3**	**65.50**		**59.6**	**59.19**

Table 3: **Dataset Coverage and Average Attribute Purity**. For LDA, we assigned a tweet to its most probable topic if the probability was ≥ 0.5.

Effects of Consecutive Iterations In Fig. 1 we observed different behaviors in subsequent iterations with respect to coverage and performance. To further inspect this phenomenon, we looked at the tweets that shifted predictions between the first and second iterations. Fig. 3 shows this analysis for Immigration. Here, we find that a considerable number of tweets that were assigned to a theme in the first iteration, were unmatched (e.g. moved to the *"Unknown"*) in the second iteration. This behavior explains the decrease in coverage. Upon closer inspection, we found that the majority of these unmatched tweets corresponded to assignments that were in the last and second to last intervals with respect to their similarity to the theme embedding. We also observed a non-trivial movement from the unknown to the new themes (shown in red), as well as some shifts between old themes and new themes that seem reasonable. For example, 1.2% of the total tweets moved from *"Role of Western Countries"* to *"Country of Immigrants"*, 1% moved from *Academic Discussions* to *Activism*, and close to 3% of tweets moved from *Trump Policy* and *UK Policy* to *Criticize Anti Immigrant Rhetoric*. This behavior, coupled with the increase in performance observed, suggests that as new themes are added, tweets move to a closer fit.

In App. A.7 we include the matrix of shifted predictions for Covid, as well as the details of distribution of the unmatched tweets with respect to their semantic similarity to the theme embeddings. For Covid, we observe that the increase in coverage is attributed to the addition of the *"Vax Efforts Progression"* theme, which encompasses all mentions to vaccine development and roll-out. Otherwise, a similar shifting behavior can be appreciated.

Consistency between Different Expert Groups: To study the subjectivity of experts and its impact on the resulting themes, we performed two parallel studies on the Covid corpus. For each study, a different group of experts performed

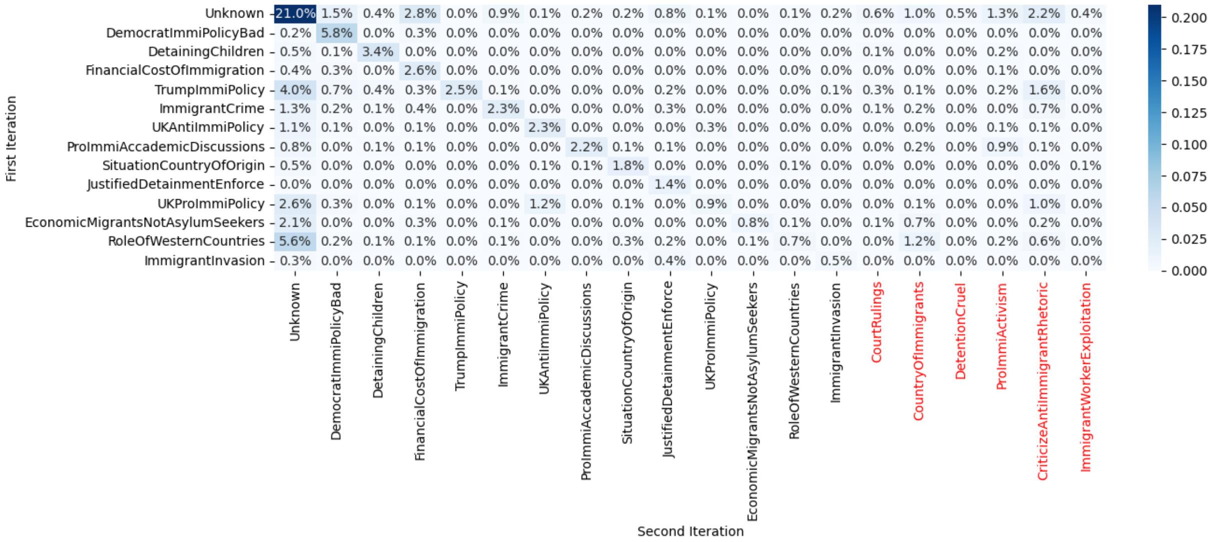


Figure 3: **Shifting predictions for Immigration**. Themes added during second iteration are shown in red, and values are normalized over the full population.

Iter.	Metric	Group 1	Group 2
1	Num Themes	9	8
	Coverage	54.30	61.80
	Stance Purity	83.18	87.43
	Moral Found. Purity	56.75	65.52
2	Num Themes	16	14
	Coverage	84.30	85.90
	Stance Purity	80.12	84.31
	Moral Found. Purity	50.88	52.17

Table 4: Two Different Groups of Experts on **Covid**

two rounds of interaction following the protocol outlined on Sec. 2.2. The side-by-side comparisons of the two studies can be observed in Tab. 4

We find that the second group of experts is able to obtain higher coverage and higher attribute purity with a slightly reduced number of themes. To further inspect this phenomenon, as well as the similarities and differences between the two sets of themes, we plot the overlap coefficients between the theme-to-tweet mappings in Fig. 4. We use the Szymkiewicz–Simpson coefficient, which measures the overlap between two finite sets and is defined as: $overlap(X, Y) = \frac{|X \cap Y|}{min(|X|, |Y|)}$ In cases where we observe high overlap between the two groups, we find that there is essentially a word-for-word match between the two discovered themes. For example, *Vax Lessens Symptoms*, which was surprisingly named the same by the two groups, as well as *Vax Availability* vs. *Vax Appointments*, *Got The Vax* vs. *I Got My Vax*, and *Vax Side Effects* vs. *Post Vax Symptoms*. In other cases, we find that different groups came up with themes that have some conceptual (and literal) overlap, but that span different sub-segments of the data. For example, we see that the theme *Reasons the US Lags On Vax* defined by the second group, has overlap with different related themes in the first group, such as: *Gov. Bad Policies*, *Vax Efforts Progression*, and *Unjustified Fear of Vax*. Similarly, while the first group defined a single theme *Vax Personal Choice*, the first group attempted to break down references to personal choices between those direclty related to taking the vaccine (*Free Choice Vax*), and those that use the vaccine as analogies for other topics, like abortion (*Free Choice Other*).

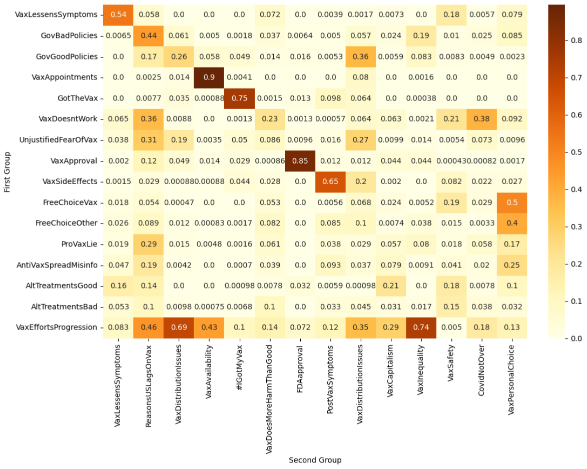


Figure 4: Theme Overlap Coefficient Heatmap between Different Groups of Experts

While some themes are clearly present in the data and identified by the two groups, we see that subjective decisions can influence the results. The first group was inclined to finer grained themes (with the exception of *Vax Efforts Progression*), while the second group seemed to prefer more general themes. In future work, we would like to study how the variations observed with our approach compare to the variations encountered when experts follow fully manual procedures, as well the impact of the crowd vs. experts working alone.

Abstract Themes vs. Word-level Topics: To get more insight into the differences between topics based on word distributions and our themes, we looked at the overlap coefficients between topics obtained using LDA and our themes. Fig. 24 shows the coefficients for Immigration. While some overlap exists, the coefficients are never too high (a max. of 0.35). One interesting finding is that most of our themes span multiple related topics. For example, we find that *Trump Policy* is has similar overlap with *undocumented_ice_workers_trump*, *migrants_migrant_trump_border*, and *children_parent_kids_trump*. While all of these topics discuss Trump policies, they make reference to different aspects: workers, the border and families. This supports our hypothesis that our themes are more abstract in nature, and that capture conceptual similarities beyond word distributions. Overlap coefficients for Covid and for subsequent iterations can be seen in App. A.8.

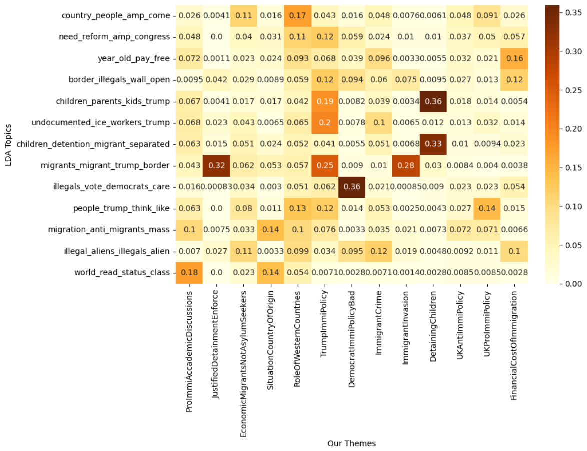


Figure 5: Overlap Coefficients between LDA Topics and our Themes (First Iteration for Immigration).

4 Related Work

This paper suggests a novel approach for identifying themes emerging from a document collection. The notion of a theme is strongly related to topic models (Blei et al., 2003). However unlike latent topics that are defined as a word distributions, our goal is to provide a richer representation by connecting the themes to general concepts that help explain them, such as moral foundations theory (Haidt and Graham, 2007; Amin et al., 2017; Chan, 2021; Roy et al., 2021) and framing theory (Entman, 1993; Chong and Druckman, 2007; Morstatter et al., 2018).

Our work is conceptually similar to several recent works that characterize themes and issue-specific frames in data, either by manually developing a codebook and annotating data according to it (Boydstun et al., 2014; Mendelsohn et al., 2021), or by using data-driven methods (Demszky et al., 2019; Roy and Goldwasser, 2021). Unlike these approaches our work relies on interleaved human-machine interaction rounds, in which humans can identify and explain themes from a set of candidates suggested by the model, as well as diagnose and adapt the model's ability to recognize these themes in documents. This work is part of a growing trend in NLP, which studies how human-machine collaboration can help improve language learning (Wang et al., 2021). In that space, two lines of works are most similar to ours. Interactive topic models (Hu et al., 2011; Lund et al., 2017; Smith et al., 2018) allow humans to adapt the topics using lexical information. Open Framing (Bhatia et al., 2021) allows humans to identify and name frames based on the output of topic models, but lacks our model's ability for sustained interactions that shape the theme space and the explanatory power of our neuro-symbolic representation.

5 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 an exhaustive evaluation of our framework using two case studies and different groups of experts. Additionally, we 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.

References

Avnika B Amin, Robert A Bednarczyk, Cara E Ray, Kala J Melchiori, Jesse Graham, Jeffrey R Huntsinger, and Saad B Omer. 2017. Association of moral values with vaccine hesitancy. *Nature Human Behaviour*, 1(12):873–880.

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

Rodney Benson. 2013. *Shaping Immigration News: A French-American Comparison*. Communication, Society and Politics. Cambridge University Press.

Vibhu Bhatia, Vidya Prasad Akavoor, Sejin Paik, Lei Guo, Mona Jalal, Alyssa Smith, David Assefa Tofu, Edward Edberg Halim, Yimeng Sun, Margrit Betke, Prakash Ishwar, and Derry Tanti Wijaya. 2021. OpenFraming: Open-sourced tool for computational framing analysis of multilingual data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 242–250, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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*.

Amber Boydstun, Dallas Card, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2014. Tracking the development of media frames within and across policy issues.

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

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *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)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Eugene Y Chan. 2021. Moral foundations underlying behavioral compliance during the covid-19 pandemic. *Personality and individual differences*, 171:110463.

Dennis Chong and James N Druckman. 2007. Framing theory. *Annu. Rev. Polit. Sci.*, 10:103–126.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *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)*, pages 2970–3005.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Robert M Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of communication*, 43(4):51–58.

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

Jan Fredrik Hovden and Hilmar Mjelde. 2019. Increasingly controversial, cultural, and political: The immigration debate in scandinavian newspapers 1970–2016. *Javnost - The Public*, 26(2):138–157.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. In *Advances in Neural Information Processing Systems*, volume 34, pages 2018–2033. Curran Associates, Inc.

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.

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

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

Ralph H. Johnson. 1988. Gilbert harman change in view: Principles of reasoning (cambridge, ma: Mit press 1986). pp. ix 147. *Canadian Journal of Philosophy*, 18(1):163–178.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality.

In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden. Association for Computational Linguistics.

Claire Lauer, Eva Brumberger, and Aaron Beveridge. 2018. Hand collecting and coding versus data-driven methods in technical and professional communication research. *IEEE Transactions on Professional Communication*, 61(4):389–408.

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.

Fred Morstatter, Liang Wu, Uraz Yavanoglu, Stephen R. Corman, and Huan Liu. 2018. Identifying framing bias in online news. *Trans. Soc. Comput.*, 1(2):5:1–5:18.

Maria Leonor Pacheco and Dan Goldwasser. 2021. Modeling content and context with deep relational learning. *Transactions of the Association for Computational Linguistics*, 9:100–119.

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.

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

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.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.

T. Rogers and James L. McClelland. 2004. Semantic cognition: A parallel distributed processing approach.

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

Shamik Roy and Dan Goldwasser. 2021. Analysis of nuanced stances and sentiment towards entities of US politicians through the lens of moral foundation theory. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 1–13, Online. Association for Computational Linguistics.

Shamik Roy, Maria Leonor Pacheco, and Dan Goldwasser. 2021. Identifying morality frames in political tweets using relational learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9939–9958, Online and Punta Cana, Dominican Republic. 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.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961, Jeju Island, Korea. Association for Computational Linguistics.

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

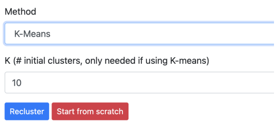
Zijie J. Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. 2021. Putting humans in the natural language processing loop: A survey. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 47–52, Online. Association for Computational Linguistics.

Hongteng Xu, Wenlin Wang, Wei Liu, and Lawrence Carin. 2018. Distilled wasserstein learning for word embedding and topic modeling. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

A Appendix

A.1 Tool Screenshots

A.1.1 Exploratory Operations



Method: K-Means

K (# initial clusters, only needed if using K-means): 10

Figure 6: Cluster Instances

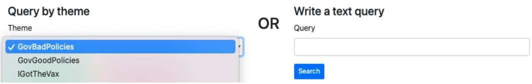
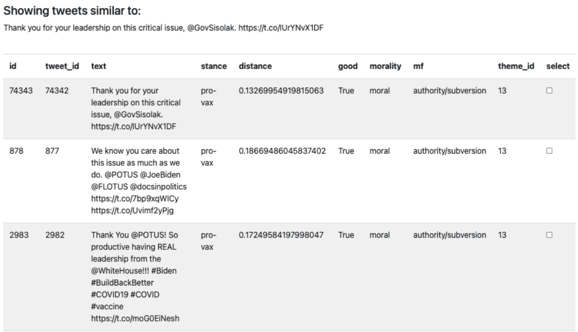


Figure 7: Text-based Queries

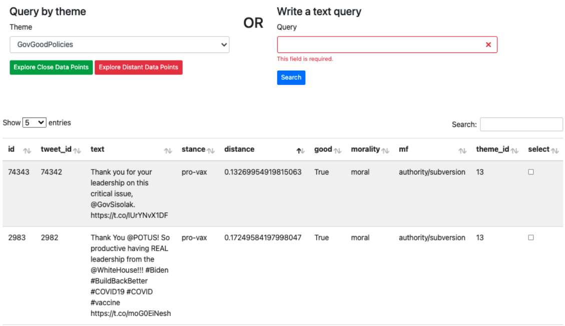


Showing tweets similar to:

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 8: Finding Similar Tweets



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 9: Listing Arguments and Examples

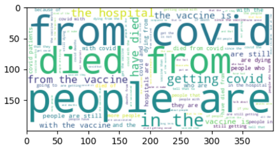
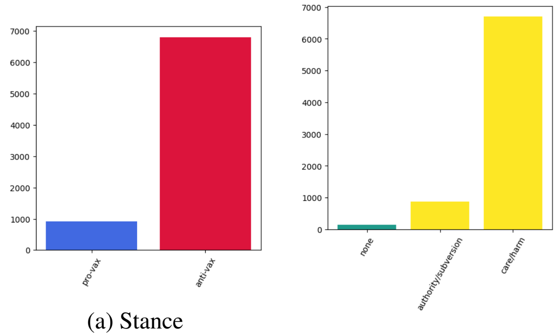


Figure 10: Visualizing Local Explanations: Word Cloud Example for *The Vaccine Doesn't Work*



(a) Stance

(b) Moral Foundation

Figure 12: Visualizing Local Explanations: Attribute Distribution for *The Vaccine Doesn't Work*

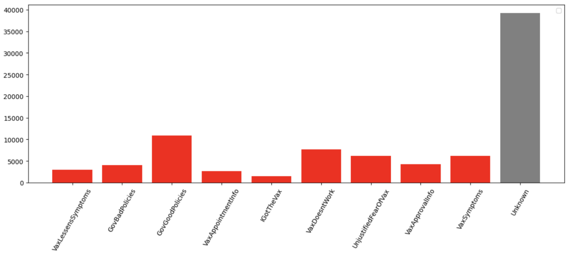
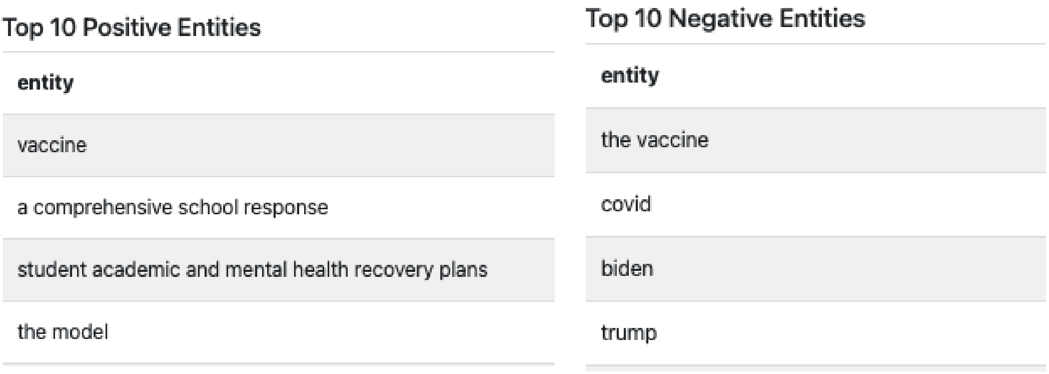


Figure 13: Visualizing Global Explanations: Theme Distribution



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

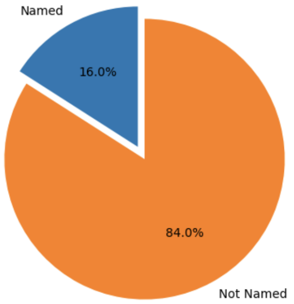


Figure 14: Visualizing Global Explanations: Coverage

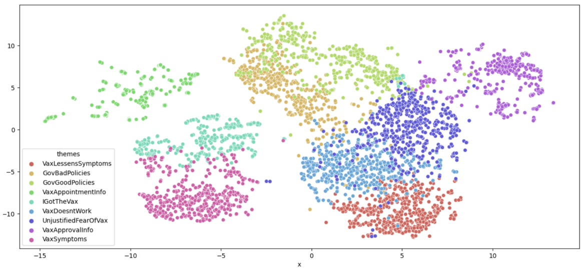
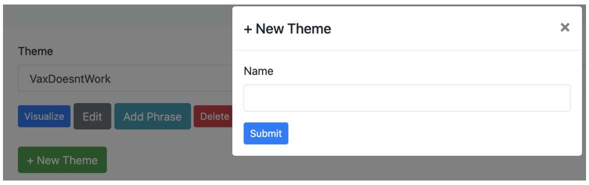


Figure 15: Visualizing Global Explanations: 2D t-SNE

A.1.2 Intervention Operations

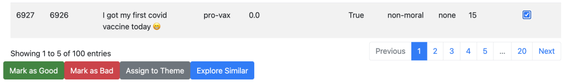


+ New Theme

Name

Submit

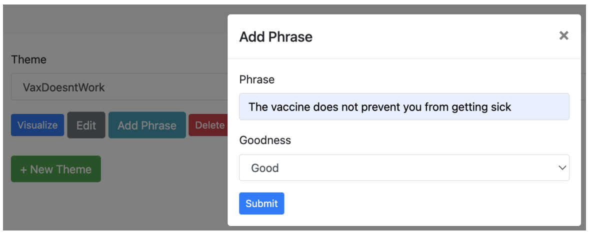
Figure 16: Adding New Themes



6927 6926 I got my first covid vaccine today 😁 pro-vax 0.0 True non-moral none 15 ☑

Showing 1 to 5 of 100 entries

Figure 17: Marking Instances as *Good*



Add Phrase

Phrase

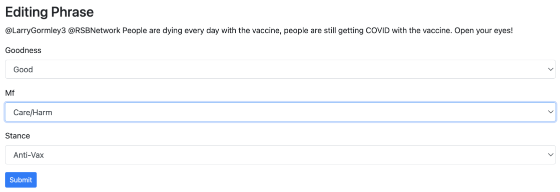
The vaccine does not prevent you from getting sick

Goodness

Good

Submit

Figure 18: Adding *Good* Examples



Editing Phrase

@LarryGormley3 @RSBNetwork People are dying every day with the vaccine, people are still getting COVID with the vaccine. Open your eyes!

Goodness

Good

Mf

Care/Harm

Stance

Anti-Vax

Submit

Figure 19: Correcting Attributes - Stances and Moral Foundations

A.2 Interactive Sessions for Covid: First Group of Experts

Table 5 and 6 outline the patterns discovered by the the first group of experts on the first a second iteration, respectively.

A.3 Interactive Sessions for Covid: Second Group of Experts

Table 7 and 8 outline the patterns discovered by the second group of experts on the first a second iteration, respectively.

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 5: **First Iteration**: 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 6: **Second Iteration**: Patterns Identified in Subsequent Clusters and Resulting Themes

Cluster	Experts Rationale	New Named Themes
K-Means 0	People asking people to get vaccinated. Some skeptical but acknowledge it reduces symptoms. It works but it has limitations. More specifically, it lessens the symptoms.	VaxLessensSymptoms
K-Means 1	Republicans have hurt the vax rate in the US. Finding someone (or some party) to blame. Politicians are hurting people with policy. Vaccine in the US is behind, trying to explain why	ReasonsUSLagsOnVax
K-Means 2	A lot of them are just replies. Cluster is for links and usernames.	-
K-Means 3	Availability and distribution of the vaccine. How stances of people in different states affect it. Vaccine distribution issues due to local policy.	VaxDistributionIssuesDueToLocalPolicy
K-Means 4	Clear cluster. Vaccine info, availability info.	VaxAvailabilityInfo
K-Means 5	Testimonials, #IGotMyVax	#IGotMyVax
K-Means 6	Some themes match the vaccine lessens symptoms. Other theme: no need to get the vaccine, it doesn't work. Vaccine does more harm than good.	VaxDoesMoreHarmThanGood
K-Means 7	Same as K-means 5	-
K-Means 8	About covid vaccine updates. FDA approval. In other cases it depends on the content on the link. So you can't really tell.	FDAApproval
K-Means 9	Obvious. Vaccine symptoms, vaccine effects. Post vaccination symptoms.	PostVaxSymptoms

Table 7: **Second Group's First Iteration**: Patterns Identified in Initial Clusters and Resulting Themes

Cluster	Experts Rationale	New Named Themes
K-Means 0	Links and promotions	-
K-Means 1	Looks like previous theme IGotMyVax, assign them.	-
K-Means 2	Very short tweets with links, and no context. Could be availability but not sure. Decided against adding theme	-
K-Means 3	Two themes observed. One old one, regarding VaxAvailabilityInfo. One new one, getting vaccines is difficult. Not related to local policy. **Decided against merging with previous theme**	VaxDistributionIssues
K-Means 4	A lot of talk about skepticism regarding the vaccine. Some good matches to previous MoreHarmThanGood, assign them. Mentions to profiting from the vaccine. **Look for similar instances to mentions of profits** **Text look up for "vaccine getting rich"** Mentions to redlining, implications of inequality **Text look up for "vaccine inequality"** Lots of mentions to racial and monetary inequalities in access to vaccine	VaxCapitalism VaxInequality
K-Means 5	Both PostVaxSymptoms and IGotMyVax examples, assign them.	-
K-Means 6	Mentions to vaccine safety. Weighting the safety/risks of the vaccine	VaxSafety
K-Means 7	A lot of discussion about the pandemic not being over Discussion on whether to open back up or not	CovidNotOver
K-Means 8	Repetitions, IGotMyVax. Assign them.	-
K-Means 9	Mentions to mandates. The vaccine should be a personal choice, mandates should not be there. Different reasons: personal choice, no proof of whether it works. For no proof, assign to previous MoreHarmThanGood	VaxPersonalChoice

Table 8: **Second Group's Second Iteration**: Patterns Identified in Subsequent Clusters and Resulting Themes

A.4 Interactive Sessions for Immigration

Table 9 and 10 outline the patterns discovered by the experts for immigration.

A.5 Topic Modeling Details

To obtain LDA topics, we use the Gensim implementation (Rehurek and 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.

A.6 Fine-Grained Results

The confusion matrix for Immigration can be seen in Fig. 20. Distribution of errors that do not match any existing theme, according to their similarity interval can be seen in Fig 21.

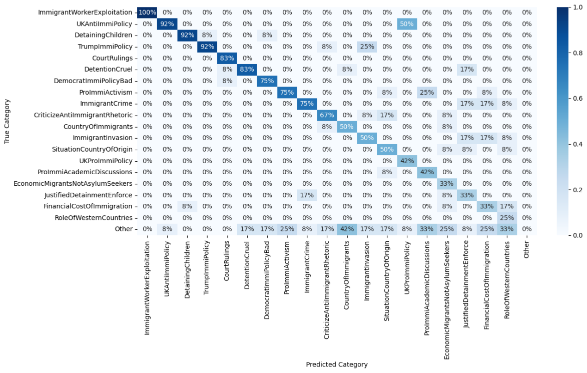
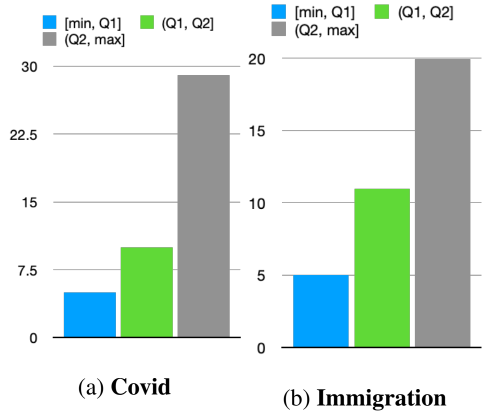


Figure 20: **Confusion matrix of Immigration themes after second iteration**. Values are normalized over the predicted themes (columns), and sorted from most accurate to least accurate.



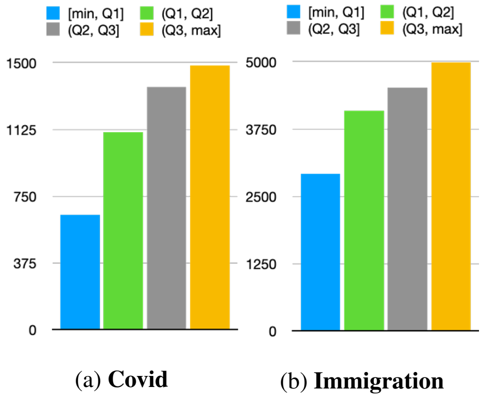
(a) **Covid**

(b) **Immigration**

Figure 21: Tweets that Do Not Match Current Set of Themes (True Category is "Other") at Different Intervals

A.7 Shifting Predictions between Iterations

Heatmaps of shifting predictions for Covid can be seen in Fig. 22. The distribution of the unmatched predictions for both Covid and Immigration, according to their similarity intervals can be seen in Fig. 23. Additionally, some examples of shifting predictions for the two themes with the most movement for the Immigration case can be seen in Tabs. 11 and 12.



(a) **Covid**

(b) **Immigration**

Figure 23: Unmatched Predictions (Shifting from Named Theme to Unknown) at Different Intervals

A.8 LDA vs. our Themes

Overlap coefficient heatmaps between LDA topics and our themes for Covid can be seen in **??**. Similarly, they can be seen for both Covid and Immigration in Fig. 25.

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 9: **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 10: **First Iteration Immigration**: Patterns Identified in Initial Clusters and Resulting Themes

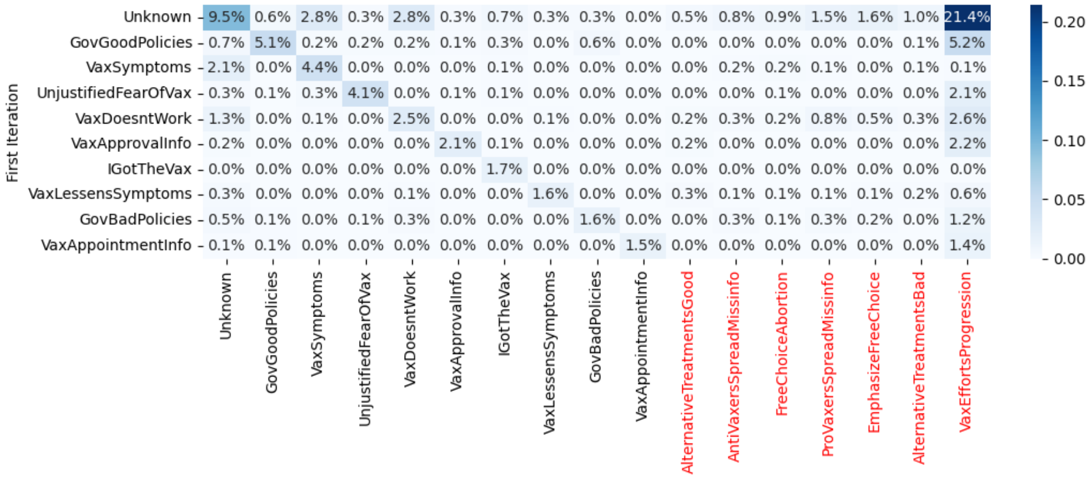


Figure 22: **Shifting predictions for Covid**. Themes added during second iteration are shown in red, and values are normalized over the full population.

Distance to Centroid	Example Tweets Kept on *Role of Western Countries*	Example Tweets Shifted to *Unknown*
0.27	The U.S. Helped Destabilize Honduras. Now Honduran Migrants Are Fleeing Political and Economic Crisis	Interesting that your problem is with "migrants", where the U.S. has issues with illegal aliens, that even our legal migrants wish to be rid of.
0.29	These people are fleeing their countries DIRECTLY because of U.S. ForeignPolicy. If you don't like refugees. Don't create 'em.	The root causes of migration aren't being addressed ASAP, as they must be. The governments are all busy talking about stopping the consequences without concrete plans to solve the causes.
0.30	Don't want migrants? Stop blowing their countries to pieces	What's missing in the US corporate news on migrants is the way American "aid" is used to overturn democracies, prop up strongmen and terrify the opposition.

Table 11: *Role of Western Countries*: Examples of tweets kept on theme (Left) and shifted to unknown (Right) between the first and second iteration. On Right are the tweets closest to the theme centroid that shifted to *Unknown*. On Left are tweets that did ***not*** shift, but have the same distance.

Distance to Centroid	Example Tweets Kept on *Trump Immigration Policy*	Example Tweets Shifted to *Unknown*
0.24	Racist realDonaldTrump wastes our tax money on locking up little kids in #TrumpConcentrationCamps and steals from our military to waste money on his #ReElectiomHateWall and spends little on anything else.	The anti-migrant cruelty of the Trump Admin knows no bounds. This targeting of migrant families is meant to induce fear and doesnt address our broken immigration system. We should be working to make our immigration system more humane, not dangerous and cruel.
0.25	Trump promises immigration crackdown ahead of U.S. election	This is unlawful and is directed at mothers with their children! He had no remorse for separating immigrants earlier, now he's threatening their lives! It's heart wrenching, but Trumpf has no heart! He's void of feeling empathy! Read they are in prison camps? WH ignoring cries
0.26	Trump to end asylum protections for most Central American migrants at US-Mexico border	BBC News - Daca Dreamers: Trump vents anger on immigrant programme

Table 12: *Trump Immigration Policy*: Examples of tweets kept on theme (Left) and shifted to unknown (Right) between the first and second iteration. On Right are the tweets closest to the theme centroid that shifted to *Unknown*. On Left are tweets that did ***not*** shift, but have the same distance.

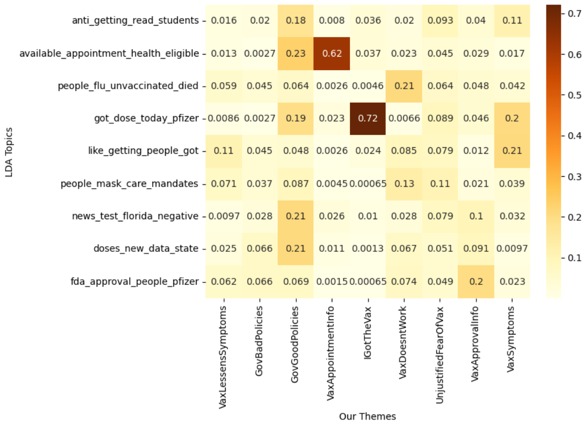
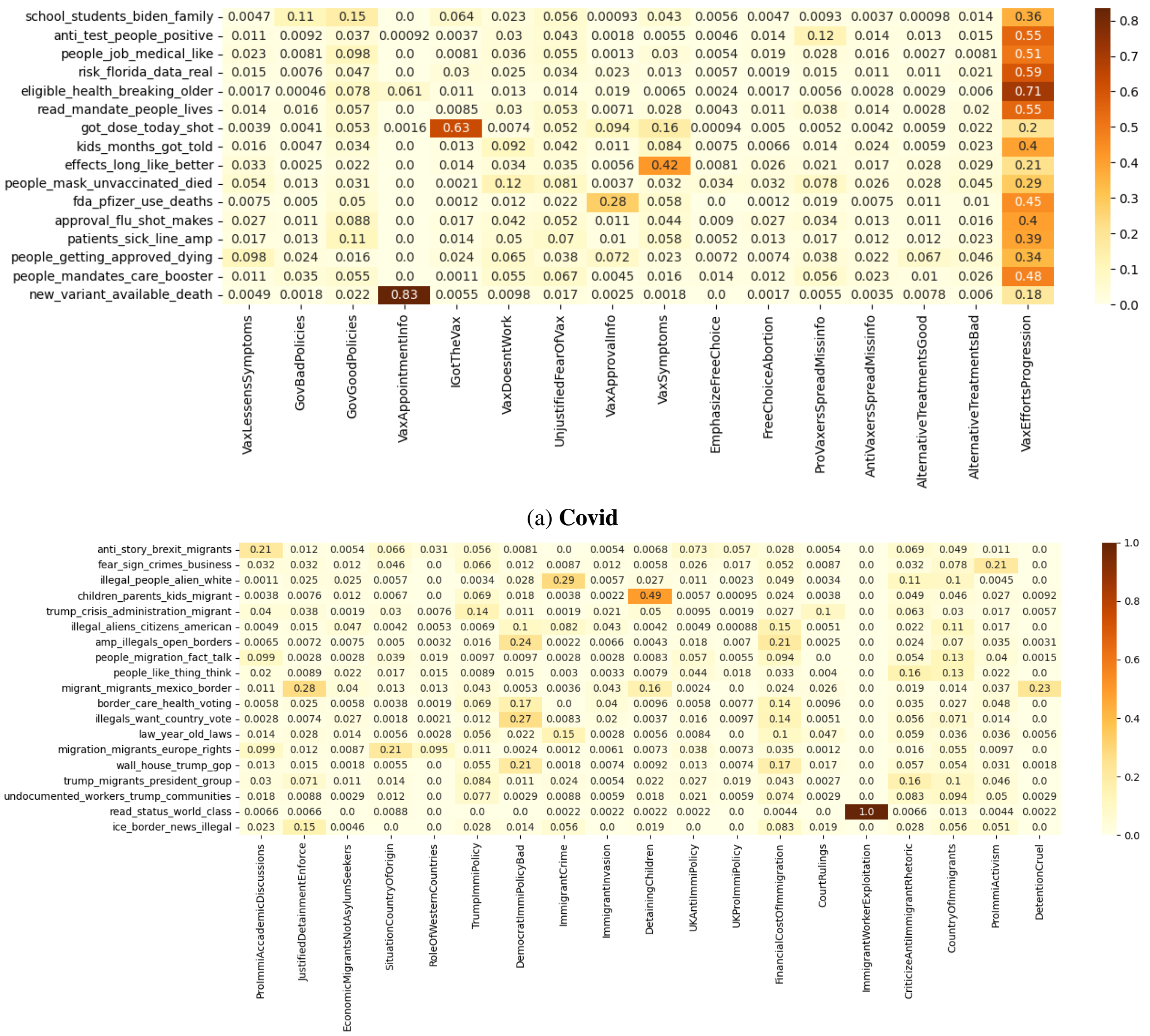


Figure 24: Overlap Coefficients between LDA Topics and our Themes (First Iteration for Covid).



(a) **Covid**

(b) **Immigration**

Figure 25: Overlap Coefficients between LDA Topics and our Themes (First Iteration). LDA Topics are represented by their 4 most prominent words.