A Holistic Framework for Analyzing the COVID-19 Vaccine Debate

Anonymous ACL submission

Abstract

The Covid-19 pandemic has led to infodemic of low quality information leading to poor health decisions. Combating the outcomes of this infodemic is not only a question of identifying false claims, but also reasoning about the decisions individuals make. In this work we propose a holistic analysis framework connecting stance and reason analysis and fine-grained entity level moral sentiment analysis. We study how to model the dependencies between the different level of analysis and incorporate human insights into the learning process. Experiments show that our framework provides reliable predictions even in the low-supervision settings.

1 Introduction

One of the unfortunate side-effects of the Covid-19 pandemic is a global infodemic flooding social media with low quality and polarizing information about the pandemic, influencing its perception and risks associated with it (Tagliabue et al., 2020). As studies have shown (Montagni et al., 2021), these influences have clear real-world implications, in terms of public acceptance of treatment options, vaccination and prevention measures.

Most computational approaches tackling the Covid-19 infodemic view it a misinformation detection problem, i.e., identifying false claims and analyzing reactions to them on social media (Hossain et al., 2020; Alam et al., 2021; Weinzierl et al., 2021). This approach, while definitely a necessary component in fighting the infodemic, does not provide policy makers and health-professionals with much needed information, characterizing the reasons and attitudes that underlie the health and well-being choices individuals make.

Our goal in this paper is to suggest a holistic analysis framework, providing multiple interconnected views of the opinions expressed in text. We specifically focus on a timely topic, attitudes explaining vaccination hesitancy. Figure 1 describes an example of our framework. Our analysis identifies the *stance* expressed in the post (`anti-vaccination`) and the *reason* for it (`distrust of government`). Given the ideologically polarized climate of social media discussion on this topic, we also aim to characterize the moral attitudes expressed in the text (`oppression`), and how different entities mentioned in it are perceived (`"Biden, Government" are oppressing, "citizens, us" are oppressed`). When constructing this framework we tackled three key challenges.

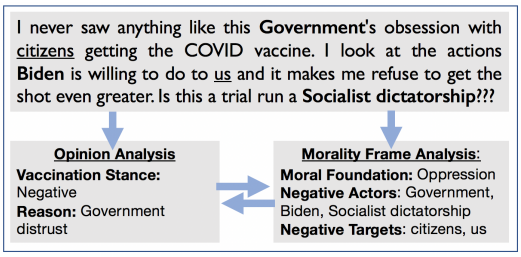


Figure 1: Holistic Analysis Framework of Social Media Posts, Connecting entity-level Moral Perspectives, Stance and Arguments Justifying it.

1. How should these analysis dimensions be operationalized? While stance prediction is an established NLP task, constructing the space of possible arguments justifying stances on a given topic, and their identification in text, are still open challenges. *We take a human-in-the-loop approach to both problems*. We begin by defining a seed set of relevant arguments based on data-driven studies (Weinzierl et al., 2021; Sowa et al., 2021), each reason defined by a single exemplar sentence. In a sequence of interactions, we use a pre-trained textual-inference model to identify paraphrases in a large collection of Covid-19 vaccination tweets, and present a visualization of the results to humans, which perform error analysis and based on it either add more sentences to help characterize the reason better, or add and characterize additional reasons, based on examples retrieved from the large corpus.

We explain this process in detail in Sec. 2

Our morality analysis is motivated by social science studies (Pagliaro et al., 2021; Díaz and Cova, 2021; Chan, 2021) that demonstrate the connection between moral foundation preferences (Haidt and Graham, 2007; Graham et al., 2009) and Covid-related health choices, for example showing that the endorsement of *fairness* and *care* moral foundations is correlated with trust in science. To account for fine-grained patterns, we adapt the recently proposed morality-frame formalism (Roy et al., 2021) that identifies moral roles associated with moral foundation expressions in text. These roles correspond to actor/target roles (similar to agent/patient) and positive or negative polarity, which should be understood in the context of a specific moral foundation. In Fig. 1 *"Biden"* is the negative actor in the context of Oppression, making him the oppressor. We explain this formalism in Sec. 3.

2. How should the dependencies between these dimensions be captured and utilized? The combination of stance, reason and moral attitudes provides a powerful source of information, allowing us to capture the moral attitudes expressed in the context of different stances and their reasons. These connections can also be utilized to help build expectations about likely attitudes in the context of each stance. As a motivating example, consider the reason "`distrust in government`", which can be associated with the "`oppression`" moral foundation, however only when its actor is an entity related to government functions (rather than oppression of Covid-19 illness). We model these expectation as a probabilistic inference process (Pacheco and Goldwasser, 2021), by incorporating consistency constraints over the judgements made by our model, and predicting jointly the most likely analysis, consisting of all analysis dimensions. The full model, described using a declarative modeling language, is provided in Section 5.

3. How can text analysis models be adapted to this highly dynamic domain, without extensive and costly manual annotation? While our analysis in this paper focuses on a specific issue, vaccination hesitancy, we believe that our analysis framework should be easily adaptable to new issues. Relying on human insight to characterize and operationalize stance and reason identification is one aspect, that characterizes *issue-specific* considerations. Moral Foundation Theory, by its definition abstracts over specific debate topics, and offers a general account for human morality. However, from a practical perspective, models for predicting these highly abstract concepts are trained on data specific to a debate topic and might not generalize well. Instead of retraining the model from scratch, we hypothesize that given an initial model constructed using out-of-domain data, and a small amount of in-domain labeled data, we can obtain acceptable performance by modeling the interaction between reasons, stances and moral foundations. We study these settings, along with the fully supervised setting in Sec. 6.

2 Opinion Analysis

To analyze opinions about the COVID-19 vaccine, we model the vaccination stance expressed in each tweet (i.e. pro-vaccine, anti-vaccine, neutral) and the underlying reason behind such stance. For example, in Fig. 1 the tweet expresses an anti-vaccine stance, and mentions their distrust of the Biden administration as the reason to take this stance.

There are three main challenges involved in this analysis: 1) predicting the stance, 2) constructing the space of possible reasons, and 3) mapping tweets to the relevant reasons. Stance prediction is an established NLP classification task (Glandt et al., 2021). However, uncovering latent themes from text automatically remains an open challenge, traditionally approached using noisy unsupervised techniques such as topic models (Zamani et al., 2020b), or by manually identifying and annotating them in text (Hasan and Ng, 2014).

Instead, we combine computational and qualitative techniques to uncover the most frequent reasons cited for pro and anti vaccination stances. We build on previous health informatics studies that characterized the arguments made against the COVID-19 vaccine in social media (Wawrzuta et al., 2021). In this work, researchers come up with a code-book of 12 main themes, frequently used as reasons to refuse or cast doubt on the vaccine. **We propose an interactive, humans-in-the-loop protocol** to learn representations for these 12 initial reasons, ground them in data, evaluate their quality, and refine them to better capture the discussion. To do this, we build a tool to explore repeating arguments and their reasons in the COVID-19 vaccine debate. The tool consists of an interactive Google Colab notebook equipped with a custom API to query current arguments, ground them in data, and visualize them. To initialize the system,

show_reasons() lists the current list of reasons (e.g. Government Distrust, Natural Immunity.)
show_closest_tweets(reason, K) lists the K tweets closest to a given reason, based on their embedding similarity.
wordcloud(reason) Renders a word cloud to visualize the arguments associated to a given reason, based on bigram and trigram TF-IDF features.
show_assignments(threshold) Renders a bar plot showing the assignment of tweets to reasons, based on embedding similarity. An optional threshold can be used to limit assignments.
tsne(threshold) Renders a visualization of the reason clusters in a 2D map. Threshold is optional.
silhouette_score(threshold) Measures the overlapping degree between clusters. Threshold is optional.
add_reason(reason, phrase) Adds a new reason with a phrase that characterizes it in natural language
remove_reason(reason) Removes a given reason
add_phrase(reason, phrase) Adds an additional phrase to an existing reason.

Table 1: Interactive API Operations

we use the 12 reasons suggested by Wawrzuta et al. (2021), and represent them using the one-sentence explanation provided. Our main goal is to ground these reasons in a set of approximately 85,000 unlabeled tweets about the COVID-19 vaccine (details in Sec. 4). To map tweets to reasons, we use the similarity between their SBERT embeddings (Reimers and Gurevych, 2019). The interaction is centered around the operations outlined in Tab. 1. Intuitively, the first six operations allow humans to diagnose how reasons map to text, and the last three allows them to act on the result of this diagnosis, by adding and removing reasons, and modifying the phrases characterizing each reason.

We follow a simple protocol during interaction, where three human coders use the operations above to explore the initial reasons. The coders start by looking at the global picture: the reasons distribution, the 2D visualizations (van der Maaten and Hinton, 2008) and the silhouette score (Rousseeuw, 1987). Then, they query the reasons one by one, looking at the word cloud (characterizing the distribution of short phrases over all texts assigned to the reason) and the 10 closest tweets to each reason. Following these observations, there is a discussion phase in which the coders follow a thematic analysis approach (Braun and Clarke, 2012) to uncover the overarching themes that are not covered by the current set of reasons, as well as the argumentation patterns that the method fails to identify. Then, they are allowed to add and remove reasons, as well as explanatory phrases for them in natural language. Every time a reason or phrase is added or removed, all tweets are reassigned to their closest reasons. This process was done over two one-hour sessions. The coders were NLP and Computational Social Science researchers, two female and one male, between the ages of 25 and 40.

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

Table 2: Resulting Reasons

In the first session, the coders focused on adding new reasons and removing reasons that were not prevalent in the data. For example, they noticed that the initial set of reasons contained mostly anti-vaccine arguments, and added a positive reason for each negative reason (e.g. *government distrust* ⇒ *government trust*). In addition to this, they broke down the reason "Conspiracy Theory" into specific conspiracy theories, such as *Bill Gates' micro chip*, *the vaccine contains fetal tissue*, *the vaccine makes you sterile*.They also removed infrequent reasons, such as *the swine flu vaccine*. The final set of reasons can be observed in Tab. 2

In the second session, the coders focused on identifying the argumentative patterns that were not being captured by the original reason explanations, and came up with overarching patterns to create new examples to improve the representation of the reasons. For example, in the case of the *government distrust* reason, the coders found that phrases with strong words are needed (e.g. *F the government*), examples that suggested that the government was "good at being bad" (e.g. *the government strong record of screwing things up*), and add examples with explicit negations (e.g. *the government does not work logically*). Once patterns were identified, each coder contributed a set of 2-5 examples, which were introduced to the reason representation.

In Appendix A.1, we include screenshots of the interactive notebook, and tables enumerating the full list derived patterns and phrases. To visualize the impact of interaction, we also show the overall distribution of reasons before and after interaction, and word clouds for a select set of reasons. The methodology and tool we developed are broadly applicable for diagnosing NLP models. *The tool and its documentation will be made public.*

CARE/HARM: Underlies virtues of kindness, gentleness, and nurturance.
FAIRNESS/CHEATING: Generates ideas of justice, rights, and autonomy.
LOYALTY/BETRAYAL: Underlies virtues of patriotism and self-sacrifice for the group. It is active anytime people feel that it's "one for all, and all for one."
AUTHORITY/SUBVERSION: Underlies virtues of leadership and followership, including deference to legitimate authority and respect for traditions.
PURITY/DEGRADATION: Underlies religious notions of striving to live in an elevated, less carnal, more noble way. It underlies the widespread idea that the body is a temple which can be desecrated by immoral activities and contaminants.
LIBERTY/OPPRESSION: The feelings of reactance and resentment people feel toward those who dominate them and restrict their liberty.

Table 3: Moral Foundations

3 Morality Frame Analysis

Moral Foundations Theory (Haidt and Graham, 2007) suggests that there are at least six basic foundations that account for the similarities and recurrent themes in morality across cultures, each with a positive and negative polarity (See Tab. 3).

To analyze moral perspectives in tweets, we build on the definition of morality frames proposed by Roy et al. (2021), where moral foundations are regarded as frame predicates, and associated with positive and negative entity roles.

While Roy et al. (2021) defined different roles types for each moral foundation (e.g. *entity causing harm*, *entity ensuring fairness*), we aggregate them into two general role types: **actor** and **target**, each with an associated polarity (positive, negative). An **actor** is a "do-er" whose actions or influence results in a positive or negative outcome for the **target** (the "do-ee"). For each moral foundation in a given tweet, we identify the "entity doing good/bad" (positive/negative actor) and "entity benefiting/suffering" (positive/negative target). For example, the statement "We are suffering from the pandemic" expresses **harm** as the moral foundation, where "pandemic" is a **negative actor**, and "we" is a **negative target** (i.e. the entity suffering from the actor's actions). There can be zero, one or multiple actors and targets in a given tweet. Entities can correspond to specific individuals or groups (e.g., I, democrats, people of a given demographic), organizations (e.g., political parties, CDC, FDA, companies), legislation or other political actions (e.g., demonstrations, petitions), disease or natural disasters (e.g., Covid, global warming), scientific or technological innovations (e.g., the vaccine, social media, the Internet), among others.

We break down the task of predicting morality frames into four classification tasks. For each tweet, our goal is to predict whether it is making moral judgement or not, and identify its prominent moral foundation. For each entity mentioned in the tweet, we predict whether it is a target or a role, and whether it has positive or negative polarity.

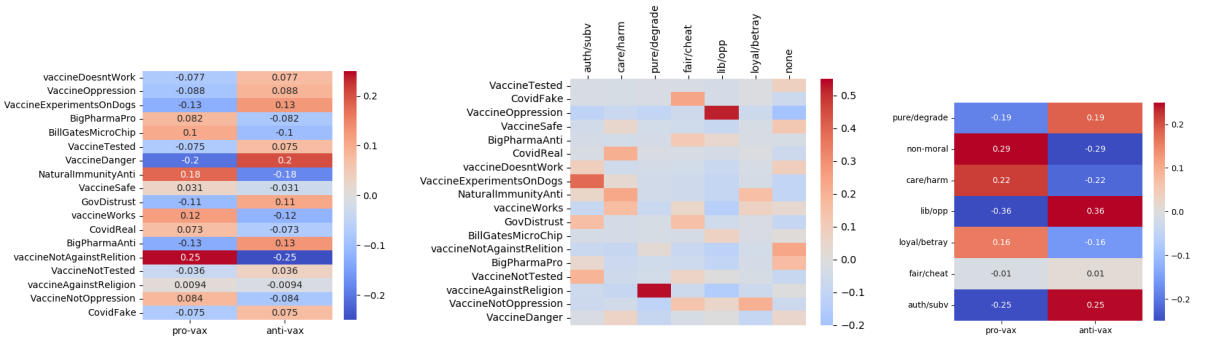
4 Data Collection and Annotation

There is no existing corpus of COVID-19 vaccine arguments annotated for morality frames and vaccination stance, so we collected and annotated our own. First, we searched for tweets between Apr. and Oct. 2021 mentioning specific keywords, such as *covid vaccine* and *vaccine mandate*. The full list of keywords, as well as the procedure to obtain them, can be seen in Appendix A.2.

Then, we created an exclusive web application for annotating our task. Moral foundation and vaccination stance labels can be annotated directly. To identify entities, annotators are able to highlight the relevant text spans, and choose its role label (i.e. positive/negative actor or target). We annotate our dataset using three in-house annotators pursuing a Ph.D. in Computer Science. We award the annotators $\$\,0.75$ per tweet and bonus ($2 * \$0.75 = \$1.5$) for completing two practice examples. Our work is IRB approved, and we follow their protocols.

To ensure quality work, we provide eight examples covering all six moral foundations and non-moral cases. Before starting the annotation task, the annotators must read the instructions, go through the examples, and annotate two practice questions. The annotation interface, examples and practice questions can be seen in Appendix A.3.

Inter-annotator agreement: We calculate the agreement among annotators using Krippendorff's α (Krippendorff, 2004), where $\alpha = 1$ suggests perfect agreement, and $\alpha = 0$ suggests chance-level agreement. We found $\alpha = 60.82$ for moral foundations, and $\alpha = 78.71$ for stance. For roles, we calculate the character by character agreement between annotations. For example, if one annotator has marked "Dr Fauci" as a target in a tweet, and another has marked "Fauci", it will be considered as an agreement on the characters "Fauci" but disagreement on "Dr". Doing this, we found $\alpha = 83.46$. When removing characters marked by all three annotators as "non-role", the agreement drops to $\alpha = 67.15$.



(a) Reasons and Vax Stance

(b) Reasons and Moral Foundations

(c) Moral Foundations and Vax Stance

Figure 2: Correlation Heatmaps

Resulting annotated dataset: We use a majority vote to get moral foundation and vaccination stance labels, and obtain 750 annotated tweets. Similarly, we define a text span to be an entity mention E, having a moral role R and polarity P, in a tweet T, if it is annotated as such by at least two annotators. Our resulting dataset contains 891 (T,E,R,P) tuples. Complete statistics can be found in Appendix A.4.[1]

To evaluate the correlation between the different dimensions of analysis, we calculate the Pearson correlation matrices and present them in Fig. 2. We can interpret reasons as distributions over moral foundations and stances (and vice-versa). This analysis provides a useful way to explain each of these dimensions. For example, we see that *care/harm* is strongly correlated with reasons such as *covid is real*, *the vaccine works*, and *natural immunity is ineffective*. Other expected trends emerge, such as *purity/degradation* being highly correlated with *vaccine against religion*. To evaluate the modeling advantage of our opinion analysis framework, we compare the reasons obtained interactively with topics extracted using LDA (Blei et al., 2003), and find that our reasons have higher correlations with both vaccination stance and moral foundations. The LDA figures can be found in Appendix A.5.

In Tab. 4 we show the top four reasons for *fairness/cheating*. We choose this moral foundation given that is evenly split among stances and is active for different reasons. We show the top two (E,R,P) tuples for each reason. We can appreciate that while this moral foundation is used by people on both sides, the reasons offered and entities used vary. On the anti-vax side, authority figures and vaccine trials are portrayed as negative actors, while women and children are portrayed as targets. On the pro-vax side, COVID and unvaccinated people are portrayed as negative actors, and the general public is portrayed as a target.

VAXNOTOPPRESSION	VAXDANGER
70% Pro-Vax (responsible people, target, neg) (un-vax people, actor, neg)	**60% Anti-Vax** (pregnant women, target, neg) (trial vax, actor, neg)
GOVDISTRUST	**VAXWORKS**
75% Anti-Vax (children, target, neg) (Fauci, actor, neg)	**75% Pro-Vax** (people, target, neg) (COVID, actor, neg)

Table 4: Top 4 reasons for **Fairness/Cheating**, and their most frequent opinions and entity roles

Unlabeled COVID-19 vaccine corpus: In addition to our annotated dataset, we collect a corpus of 85,000 tweets in English mentioning the covid vaccine, uniformly distributed between Jan. and Oct. 2021. These tweets are unlabeled, and are used to ground arguments (Sec. 2) and to augment data for indirect supervision (Sec. 5).

5 Joint Probabilistic Model

We propose a joint probabilistic model that reasons about the arguments made, their morality frames, stances, reasons, and the dependencies between them. We implement our model using DRaiL (Pacheco and Goldwasser, 2021), a declarative modeling framework for specifying deep relational models. Deep relational models combine the strengths of deep neural networks and statistical relational learning (SRL) to model a joint distribution over relational data. This hybrid modeling paradigm allow us to leverage expressive textual

[1]This dataset will be made publicly available.

encoders, and to introduce contextualizing information and model different interdependent decisions. SRL methods have proven effective to model domains with limited supervision (Johnson and Goldwasser, 2018; Subramanian et al., 2018), and approaches that combine neural nets and SRL have shown consistent performance improvements (Widmoser et al., 2021; Roy et al., 2021).

Following the conventions of statistical relational learning models, we use horn-clauses of the form $p_0 \wedge p_1 \wedge ... \wedge p_n \Rightarrow h$ to describe relational properties. Each logical rule defines a probabilistic scoring function over the relations expressed in its body and head.

Base rules/classifiers: We define three base rules to score whether a tweet $\mathtt{t_i}$ has a moral judgment, what is its prominent moral foundation $\mathtt{m}$, and what is its vaccination stance.

$$\begin{aligned} r_0 &: \mathtt{Tweet}(\mathtt{t_i}) \Rightarrow \mathtt{IsMoral}(\mathtt{t_i}) \\ r_1 &: \mathtt{Tweet}(\mathtt{t_i}) \Rightarrow \mathtt{HasMF}(\mathtt{t_i}, \mathtt{m}) \\ r_2 &: \mathtt{Tweet}(\mathtt{t_i}) \Rightarrow \mathtt{VaxStance}(\mathtt{t_i}, \mathtt{s}) \end{aligned} \tag{1}$$

To score the moral role of an entity $\mathtt{e_i}$ mentioned in tweet $\mathtt{t_i}$, we write two rules. The first one scores whether the entity $\mathtt{e_i}$ is an actor or a target, and the second one scores its polarity (positive or negative).

$$\begin{aligned} r_3 &: \mathtt{Mentions}(\mathtt{t_i}, \mathtt{e_i}) \Rightarrow \mathtt{HasRole}(\mathtt{e_i}, \mathtt{r}) \\ r_4 &: \mathtt{Mentions}(\mathtt{t_i}, \mathtt{e_i}) \Rightarrow \mathtt{EntPolarity}(\mathtt{e_i}, \mathtt{p}) \end{aligned} \tag{2}$$

Note that these rules do not express any dependencies. They function as base classifiers that map tweets and entities to their most probable labels.

Dependency between roles and moral foundations: The way an entity is portrayed in a tweet can be highly indicative of its moral foundation. For example, people are likely to mention *children* as a *negative actor* in the context of *care/harm*. To capture this, we explicitly model the dependency between an entity, its moral role, and the MF.

$$\begin{aligned} r_5 &: \mathtt{Mentions}(\mathtt{t_i}, \mathtt{e_j}) \wedge \mathtt{HasRole}(\mathtt{e_i}, \mathtt{r}) \\ &\wedge \mathtt{EntPolarity}(\mathtt{e_i}, \mathtt{p}) \Rightarrow \mathtt{HasMf}(\mathtt{t_i}, \mathtt{m}) \end{aligned} \tag{3}$$

Dependency between stances and moral foundations: As we showed in Sec. 4, there is a significant correlation between the stance of a tweet with respect to the vaccine debate, and its moral foundation. For example, people who oppose the vaccine are more likely to express the liberty/oppression MF. To capture this, we model the dependency between the stance of a tweet and its MF.

$$r_6 : \mathtt{VaxStance}(\mathtt{t_i}, \mathtt{s}) \Rightarrow \mathtt{HasMf}(\mathtt{t_i}, \mathtt{m}) \tag{4}$$

Dependency between reasons and moral foundations/stances: Explicitly modeling the dependency between repeating reasons and other decisions can help us add inductive bias into our model, potentially simplifying the task. For example, we can enforce the difference between two opposing views that use similar wording, and that could otherwise be treated similarly by a text-based model (e.g. *"natural methods of protection against the disease are better than vaccines"* vs. *'vaccines are better than natural methods of protection against the disease"*). We add two rules to capture this dependency, one between reasons and moral foundations, and one between reasons and stances.

$$\begin{aligned} r_7 &: \mathtt{Mentions}(\mathtt{t_i}, \mathtt{r}) \Rightarrow \mathtt{HasMf}(\mathtt{t_i}, \mathtt{m}) \\ r_8 &: \mathtt{Mentions}(\mathtt{t_i}, \mathtt{r}) \Rightarrow \mathtt{VaxStance}(\mathtt{t_i}, \mathtt{s}) \end{aligned} \tag{5}$$

Hard constraints: To enforce consistency between different decisions, we add two unweighted rules (or hard constraints). These rules are not associated with a scoring function and must always hold true. We enforce that, if a tweet is predicted to be moral, then it needs to also be associated to a specific moral foundation. Likewise, if a tweet is not moral, then no MF should be assigned to it.

$$\begin{aligned} c_0 &: \mathtt{IsMoral}(\mathtt{t_i}) \Rightarrow \neg\mathtt{HasMf}(\mathtt{t_i}, \mathtt{none}) \\ c_1 &: \neg\mathtt{IsMoral}(\mathtt{t_i}) \Rightarrow \mathtt{HasMf}(\mathtt{t_i}, \mathtt{none}) \end{aligned} \tag{6}$$

Whenever the tweets have the same stance, we include a constraint to enforce consistency between the polarity of different mentions of the same entity. Roy et al. (2021) showed that enforcing consistency for mentions of the same entity within a political party was beneficial. Given the polarization of the COVID-19 vaccine, we use the same rationale.

$$\begin{aligned} c_3 &: \mathtt{Mentions}(\mathtt{t_i}, \mathtt{e_i}) \wedge \mathtt{Mentions}(\mathtt{t_j}, \mathtt{e_j}) \\ &\wedge \mathtt{SameVaxStance}(\mathtt{t_i}, \mathtt{t_j}) \wedge \mathtt{EntPolarity}(\mathtt{e_i}, \mathtt{p}) \\ &\Rightarrow \mathtt{EntPolarity}(\mathtt{e_j}, \mathtt{p}) \end{aligned} \tag{7}$$

Learning and inference: The weights for each rule $w_r : p_0 \wedge p_1 \wedge ... \wedge p_n \Rightarrow h$ measure the importance of each rule in the model and can be learned from data. For example, when attempting to predict *care/harm* for a tweet $\mathtt{t_i}$, we would like the weight of rule instance $\mathtt{IsTweet}(\mathtt{t_i}) \Rightarrow \mathtt{HasMf}(\mathtt{t_i}, \mathtt{care/harm})$ to be greater than the weight of rule instance $\mathtt{IsTweet}(\mathtt{t_i}) \Rightarrow \mathtt{HasMf}(\mathtt{t_i}, \mathtt{loyalty/betrayal})$. In DRaiL, these weights are learned using neural networks with parameters θ_r. The collection of rules represents the global decision, and the solution is obtained by running a 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 and Goldwasser, 2021).

Learning with low-supervision: To learn DRaiL models in the low-supervision setting, we use an Expectation-Maximization style protocol, outlined in Algorithm 1. First, we initialize the parameters of base rules using distant supervision classifiers. For moral foundations, we use the Johnson and Goldwasser (2018) dataset and the Moral Foundation Twitter Corpus (Hoover et al., 2020). For roles, we use the Roy et al. (2021) dataset. For polarity, we combine the Roy et al. (2021) dataset with the MPQA 3.0 entity sentiment dataset (Deng and Wiebe, 2015). For vaccination stances, we annotate our 85K unlabeled tweets using a set of prominent antivax and provax hashatgs. Details about these datasets are provided in Appendix A.6.

Once the base rules have been initialized using distant supervision, we turn our attention to learning DRaiL models over the COVID-19 dataset presented in Sec. 4. We alternate between MAP inference using all rules to obtain training labels (expectation step), and training the neural nets using these labels (maximization step). We receive an optional parameter k indicating the amount of direct supervision to be used. When k is provided, $k\%$ of the annotated labels are seeded during inference.

Algorithm 1 *Low Supervision Learning Protocol*

```
Random initialization for all θ_r
for r ∈ base rules do
    θ_r ← distant supervision classifier
end for
while not converged do
    Y_gold ← DRaiL_MAP_inference(k)
    Train all rules locally using Y_gold
end while
```

6 Experimental Evaluation

The goal of our framework is to identify morality frames and opinions in tweets by modeling them jointly. In this section, we perform an exhaustive experimental analysis to evaluate the performance of our model and each of its components.

Experimental settings: In DRaiL, each rule r is associated with a neural architecture, which serves as a scoring function to obtain the rule weight w_r. We use BERT-base-uncased (Devlin et al., 2018) for all classifiers. For the rules that model dependencies (Eqs. 3, 4, 5), we concatenate the CLS token with a 1-hot vector of the symbols on the left hand side of the rule (i.e. role, sentiment, stance and reason), before passing it through a classifier. For rules that have the entity on the left-hand side (Eqs. 2, 3), we use both the tweet and the entity as an input to BERT, using the SEP token. We trained supervised models using local normalization in DRaiL, and leveraged distant supervision using protocol outlined in Alg. 1. In all cases, we used a learning rate of $2e-5$, a maximum sequence length of 100, and AdamW. In all experiments, we perform 5-fold cross-validation over the annotated dataset and report the micro-averaged results.

General results: Tab. 5 shows our general results for morality frames and vaccination stance. We evaluate our base classifiers and show the impact of modeling dependencies using DRaiL. The joint model results in a significant improvement for morality, moral foundation and vaccination stance. For entities, role and polarity remain stable. We also measure the impact of explicitly modeling reasons (Eq. 5). We find that moral foundations improve from 60.07 to 62.27 and vaccination stance improves from 67.72 to 72.53 after interaction. Full results are presented in Appendix A.7

Ablation study: We show an ablation study in Tab. 6. First, we can see how all dependencies contribute to the performance improvement, role-MF being the most impactful. We can also see that explicitly modeling morality constraints improves both the morality prediction and the MF prediction, suggesting an advantage to breaking down this decision. We observe that the stance-polarity constraint does not have a significant impact, but does not hurt performance either, suggesting that our classifiers already capture this information. Lastly, we can see that the performance for roles and polarity remains stable, potentially because these classifiers have a strong starting point.

Distant supervision: In Fig. 3 we evaluate the impact of our indirect supervision protocol by slowly augmenting the amount of direct supervision available. We can see that by leveraging out of domain-data and dependencies, we can outperform the fully supervised classifiers using 50% of the annotated labels.

MODEL	MORAL/NM		MORAL FOUND.		ACTOR/TARGET		ENT. POLARITY		VAX STANCE	
	Macro	Weighted	Macro	Weighted	Macro	Weighted	Macro	Weighted	Macro	Weighted
Random	54.96	55.36	11.07	15.15	45.57	45.72	34.63	36.69	49.16	49.23
Majority Class	37.05	43.62	8.33	23.98	34.63	36.69	46.54	58.15	35.77	39.84
Lexicon Matching	58.97	60.01	25.28	35.85	-	-	-	-	-	-
Base (distant sup.)	69.77	68.88	28.79	41.27	71.94	72.05	63.88	74.30	69.46	70.35
Base (direct sup.)	68.94	69.71	35.28	42.92	**84.71**	**84.75**	**72.92**	**84.31**	66.91	67.36
+ Joint Model	**80.53**	**81.17**	**53.29**	**62.27**	84.60	84.64	71.53	83.35	**72.06**	**72.53**

Table 5: General Results (F1 Scores). NM: Non Moral

MODEL	M/NM	MF	ACT/TAR	POLAR.
BERT	69.71	42.92	**84.75**	84.31
+RoleMF	69.71	55.54	84.64	84.13
+RoleMF+MC	79.00	57.68	84.64	84.13
+StanceMF	69.71	47.85	84.75	84.31
+StanceMF+MC	72.37	48.63	84.75	84.31
+StanceMF+MC+SPC	72.32	48.63	84.75	**84.35**
+ReasonMF	69.71	53.15	84.75	84.31
+ReasonMF+MC	72.60	53.41	84.75	84.31
+ReasonStance+SPC	69.71	42.92	84.64	83.26
+ ALL	**81.17**	**62.27**	84.64	83.26

Table 6: Ablation Study (Weighted F1). MC: Morality Constraint, SPC: Stance-Polarity Constraint

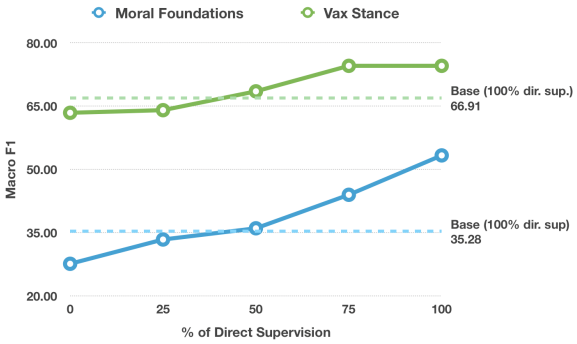


Figure 3: Performance in low-supervision settings

7 Related Work

Recent studies have noted the prevalence of rumors and misinformation in the context of the COVID-19 pandemic (Loomba et al., 2021; Shahi et al., 2021; Lazarus et al., 2021; Ahmed et al., 2020). Following this trend, several computational approaches have been proposed to detect misinformation related to COVID in news outlets and social media (Weinzierl and Harabagiu, 2021; Bang et al., 2021; Serrano et al., 2020; Al-Rakhami and Al-Amri, 2020). In this paper, we take a different approach and look at the problem of identifying opinions surrounding the COVID-19 vaccine, and explicitly modeling the rationale and moral sentiment that motivates them.

Some recent works also look at analyzing arguments about COVID and vaccine hesitancy more broadly. In most cases, they either take a traditional classification approach for predicting stances (Alliheibi et al., 2021; Lyu et al., 2021), or use topic modeling techniques to uncover trends in word usage (Skeppstedt et al., 2018; Lyu et al., 2021; Sha et al., 2020; Zamani et al., 2020a). In contrast, we propose a holistic framework that combines different methodological techniques, including human-in-the-loop mechanisms, classification with distant supervision, and deep relational learning to connect stance prediction, reason analysis and fine-grained entity moral sentiment analysis.

8 Discussion

We introduce a holistic framework for analyzing social media posts about the COVID-19 vaccine. We model morality frames and opinions jointly, and show that we can obtain competitive performance. The main limitation of our work is the size of the annotated dataset studied. Annotating for morality is a difficult and costly task, as it requires significant domain expertise. This motivates the need for methods that perform well under limited supervision, and that can leverage external and unlabeled resources. We took a first step in this direction by combining a wide range of methodological strategies. Given the amount of data generated daily about COVID, there are broader opportunities for exploiting these resources than what we explored in this paper. While we provided a preliminary analysis of the correlation between stances, reasons and morality, our current work looks at leveraging this framework to analyze opinions at scale.

We also presented a first step towards interactive exploration of opinions on social media. While we explored this technique in a very limited scenario, there is a lot of potential for using this paradigm for diagnosing NLP models and adapting to new domains. More research is required to devise better protocols and evaluation strategies for this process.

References

Wasim Ahmed, Josep Vidal-Alaball, Joseph Downing, Francesc López Seguí, et al. 2020. Covid-19 and the 5g conspiracy theory: social network analysis of twitter data. *Journal of medical internet research*, 22(5):e19458.

Mabrook S Al-Rakhami and Atif M Al-Amri. 2020. Lies kill, facts save: detecting covid-19 misinformation in twitter. *Ieee Access*, 8:155961–155970.

Firoj Alam, Fahim Dalvi, Shaden Shaar, Nadir Durrani, Hamdy Mubarak, Alex Nikolov, Giovanni Da San Martino, Ahmed Abdelali, Hassan Sajjad, Kareem Darwish, et al. 2021. Fighting the covid-19 infodemic in social media: A holistic perspective and a call to arms. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 913–922.

Fahad M Alliheibi, Abdulfattah Omar, and Nasser Al-Horais. 2021. Opinion mining of saudi responses to covid-19 vaccines on twitter. *International Journal of Advanced Computer Science and Applications*, 12(6):72–78.

Yejin Bang, Etsuko Ishii, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2021. Model generalization on covid-19 fake news detection. *arXiv preprint arXiv:2101.03841*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

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

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

Lingjia Deng and Janyce Wiebe. 2015. MPQA 3.0: An entity/event-level sentiment corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1323–1328, Denver, Colorado. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Rodrigo Díaz and Florian Cova. 2021. Reactance, morality, and disgust: The relationship between affective dispositions and compliance with official health recommendations during the covid-19 pandemic. *Cognition and Emotion*, pages 1–17.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in COVID-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1596–1611, Online. Association for Computational Linguistics.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.

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.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762, Doha, Qatar. Association for Computational Linguistics.

J. Hoover, G. Portillo-Wightman, L. Yeh, S. Havaldar, A.M. Davani, Y. Lin, B. Kennedy, M. Atari, Z. Kamel, M. Mendlen, G. Moreno, C. Park, T.E. Chang, J. Chin, C. Leong, J.Y. Leung, A. Mirinjian, and M. Dehghani. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Kristen Johnson and Dan Goldwasser. 2018. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 720–730, Melbourne, Australia. Association for Computational Linguistics.

Klaus Krippendorff. 2004. Measuring the reliability of qualitative text analysis data. *Quality and quantity*, 38:787–800.

Jeffrey V Lazarus, Scott C Ratzan, Adam Palayew, Lawrence O Gostin, Heidi J Larson, Kenneth Rabin, Spencer Kimball, and Ayman El-Mohandes. 2021. A global survey of potential acceptance of a covid-19 vaccine. *Nature medicine*, 27(2):225–228.

Sahil Loomba, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson. 2021. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature human behaviour*, 5(3):337–348.

Joanne Chen Lyu, Eileen Le Han, and Garving K Luli. 2021. Covid-19 vaccine–related discussion on twitter: topic modeling and sentiment analysis. *Journal of medical Internet research*, 23(6):e24435.

Ilaria Montagni, Kevin Ouazzani-Touhami, A Mebarki, N Texier, S Schück, Christophe Tzourio, et al. 2021. Acceptance of a covid-19 vaccine is associated with ability to detect fake news and health literacy. *Journal of public health (Oxford, England)*.

Goran Muric, Yusong Wu, and Emilio Ferrara. 2021. COVID-19 Vaccine Hesitancy on Social Media: Building a Public Twitter Dataset of Anti-vaccine Content, Vaccine Misinformation and Conspiracies.

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.

Stefano Pagliaro, Simona Sacchi, Maria Giuseppina Pacilli, Marco Brambilla, Francesca Lionetti, Karim Bettache, Mauro Bianchi, Marco Biella, Virginie Bonnot, Mihaela Boza, et al. 2021. Trust predicts covid-19 prescribed and discretionary behavioral intentions in 23 countries. *PloS one*, 16(3):e0248334.

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.

Peter Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20(1):53–65.

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.

Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. 2020. Nlp-based feature extraction for the detection of covid-19 misinformation videos on youtube. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Hao Sha, Mohammad Al Hasan, George Mohler, and P Jeffrey Brantingham. 2020. Dynamic topic modeling of the covid-19 twitter narrative among us governors and cabinet executives. *arXiv preprint arXiv:2004.11692*.

Gautam Kishore Shahi, Anne Dirkson, and Tim A Majchrzak. 2021. An exploratory study of covid-19 misinformation on twitter. *Online social networks and media*, 22:100104.

Maria Skeppstedt, Andreas Kerren, and Manfred Stede. 2018. Vaccine hesitancy in discussion forums: computer-assisted argument mining with topic models. In *Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth*, pages 366–370. IOS Press.

Paweł Sowa, Łukasz Kiszkiel, Piotr Paweł Laskowski, Maciej Alimowski, Łukasz Szczerbiński, Marlena Paniczko, Anna Moniuszko-Malinowska, and Karol Kamiński. 2021. Covid-19 vaccine hesitancy in poland—multifactorial impact trajectories. *Vaccines*, 9(8):876.

Shivashankar Subramanian, Trevor Cohn, and Timothy Baldwin. 2018. Hierarchical structured model for fine-to-coarse manifesto text analysis. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1964–1974, New Orleans, Louisiana. Association for Computational Linguistics.

Fabio Tagliabue, Luca Galassi, and Pierpaolo Mariani. 2020. The "pandemic" of disinformation in covid-19. *SN comprehensive clinical medicine*, 2(9):1287–1289.

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

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

Maxwell Weinzierl, Suellen Hopfer, and Sanda M Harabagiu. 2021. Misinformation adoption or rejection in the era of covid-19. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 787–795.

Maxwell A Weinzierl and Sanda M Harabagiu. 2021. Automatic detection of covid-19 vaccine misinformation with graph link prediction. *Journal of biomedical informatics*, 124:103955.

Manuel Widmoser, Maria Leonor Pacheco, Jean Honorio, and Dan Goldwasser. 2021. Randomized deep structured prediction for discourse-level processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1174–1184, Online. Association for Computational Linguistics.

Mohammadzaman Zamani, H Andrew Schwartz, Johannes Eichstaedt, Sharath Chandra Guntuku, Adithya Virinchipuram Ganesan, Sean Clouston, and Salvatore Giorgi. 2020a. Understanding weekly covid-19 concerns through dynamic content-specific lda topic modeling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2020, page 193. NIH Public Access.

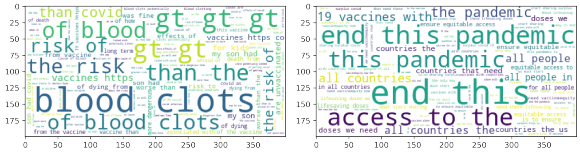
Mohammadzaman Zamani, H. Andrew Schwartz, Johannes Eichstaedt, Sharath Chandra Guntuku, Adithya Virinchipuram Ganesan, Sean Clouston, and Salvatore Giorgi. 2020b. Understanding weekly

COVID-19 concerns through dynamic content-specific LDA topic modeling. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 193–198, Online. Association for Computational Linguistics.

A Appendix

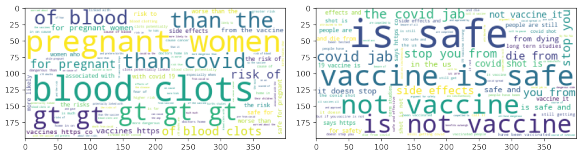
A.1 Reasons and Phrases

Tabs. 8 and 9 show the full list of phrases for anti-vax and pro-vax reasons. The interactive task interface is presented in Figs. 4 and 5. Bar plots for reason assignments before and after interaction are shown in Fig. 6.



(a) Theme: VaccineDanger (b) Theme: VaccineSafe

Figure 7: Wordclouds for reasons **before** interaction.



(a) Theme: VaccineDanger (b) Theme: VaccineSafe

Figure 8: Wordclouds for reasons **after** interaction.

A.2 Data Collection

To create the list of keywords used to collect tweets about the COVID-19 vaccine, we read multiple articles about COVID mentioning vaccination status, vaccine hesitancy, misinformation, vaccine constraints, health issues, religious sentiment and other vaccine-related debates, and made a list of repeating statements. Then, we consulted three researchers, two in Computational Social Science and one in Psychology, and constructed a list of relevant keywords that are indicative of morally charged discussions. The full list of keywords can observed in Table 10.

covid vaccine, covid vaccination, covid vaccine tyranny,
covid vaccine oppression, covid vaccine mandate, covid vaccine conspiracy,
covid vaccine anti-vax, covid vaccine religion, covid vaccine satan,
covid vaccine god, covid vaccine jesus, covid vaccine islam,
covid vaccine muslim, covid vaccine christianity, covid vaccine christian,
covid vaccine hindu, covid vaccine jews, covid vaccine catholic,
covid vaccine buddhism, covid vaccine religious, covid vaccine biden failure,
covid vaccine passport, covid vaccine loyalty, covid vaccine cheating,
covid vaccine freedom, covid vaccine betrayal, covid vaccine liberty,
covid vaccine black people, covid vaccine propaganda, covid vaccine hesitancy,
covid vaccine hesitant, covid vaccine microchip, covid vaccine bill,
covid vaccine pregnancy, covid vaccine pregnant, covid vaccine approval,
covid vaccine biden, covid vaccine fda, covid vaccine cdc,
covid vaccine fauci, covid-19 china, vaccine passport,
vaccination mandate, covid vaccine death, covid vaccine military,
experimental covid vaccine, covid vaccine authorization,
vaccine oppression, vaccine satan, covid vaccine bill gates,
covid vaccine side effect, covid vaccine adverse events

Table 10: List of the keywords for data collection.

A.3 Data Annotation Task

The steps for completing annotation in our task interface are (See Fig. 9).

1. **Select** moral foundation of the text using checkbox ☑. You can see the definition of each moral foundation by hovering mouse on them. If the tweet does not make any moral judgement, **check** ☑ "none". For this case, you don't have to highlight actor-target polarity.

2. After selecting any moral foundation other than "none", text highlighting for actor-target role with polarity will be visible below. If you select a moral foundation other than "none", you can highlight actor-target polarity.

3. **Choose** the color-coded label Positive Actor/Positive Target/Negative Actor/Negative Target to highlight the text with the color of the selected label. You can see the definition of actor-target-polarity role by hovering mouse on them.

4. **Highlight** words, phrases, or sections of the text for actor-target role with polarity of corresponding moral foundation.

5. If you made any mistake in highlighting, select "**Unhighlight**" button to unhighlight the previously highlighted text.

6. Finally, click "**Submit**" button to submit the task.

We provided eight examples (Fig. 10) covering six moral principles and non-moral cases to our

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

Table 7: Overarching argumentation patterns uncovered by coders during interaction

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

Table 8: AntiVax Themes and phrases for COVID-19 talking points. Themes that were added during interaction are shown in blue. Themes that were removed during interaction are shown in red. The original explanations/examples are presented in bold.

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

Table 9: ProVax reasons and phrases. Reasons that were added during interaction are shown in blue. Reasons that were removed during interaction are shown in red. The original explanatory phrases are presented in bold.

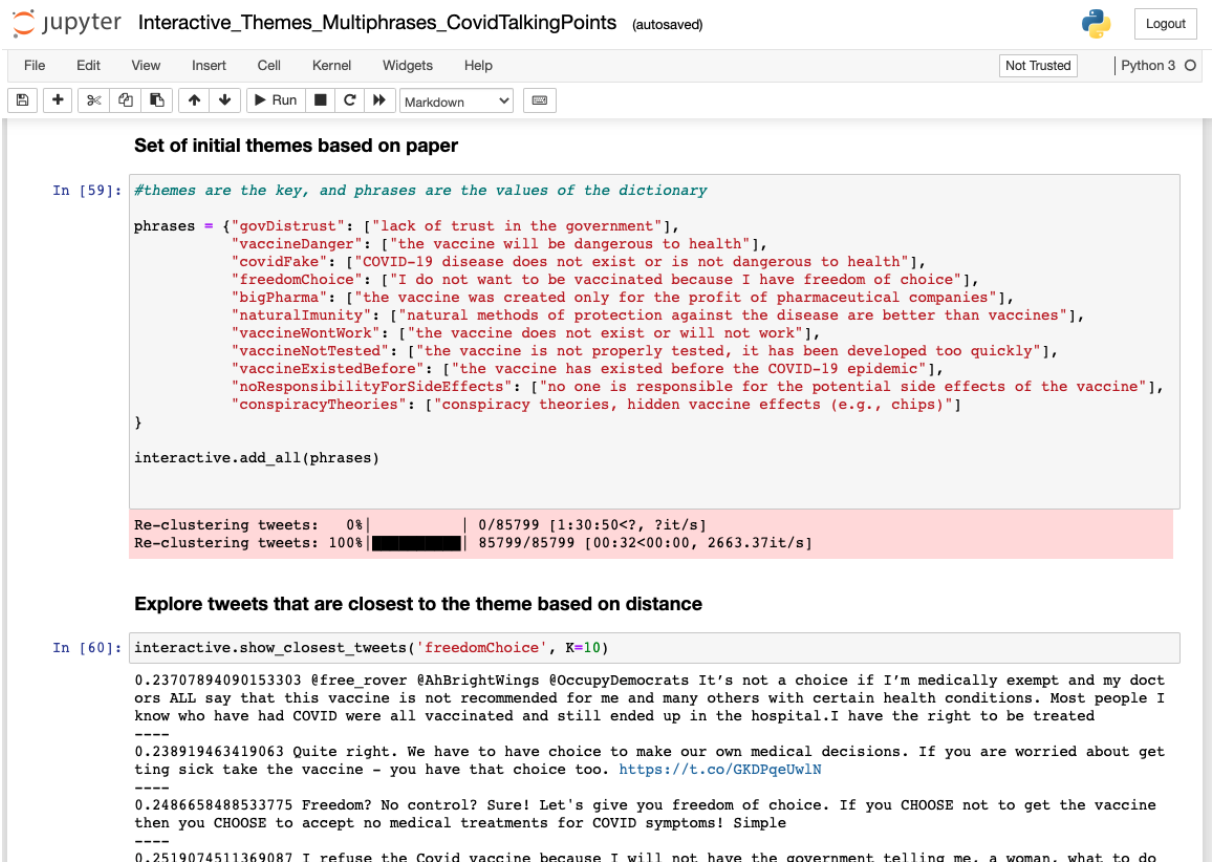


Figure 4: Interactive task interface.

annotation task interface to make it more understandable. Annotators can see the explanation behind choosing a moral foundation and actor-target polarity by clicking "**See Explanation**" button.

Annotators have to complete two practice examples before starting the real task. If they make any mistake, our practice session provides them the correct result with explanation. Fig. 11 shows the interface of one of the two practice examples.

A.4 Dataset Statistics

Full dataset statistics can be osberved in Tab. 11

Moral Foundation	Num. Tw.	Vaccination Stance			
		Pro	Anti	Neut	No Agree
Care/Harm	96	**77**	17	2	0
Fairness/Cheating	75	**33**	28	14	0
Loyalty/Betrayal	33	**26**	2	5	0
Authority/Subversion	114	26	**72**	13	3
Purity/Degradation	24	2	**22**	0	0
Liberty/Oppression	93	9	**78**	6	0
Non-moral	304	**188**	68	44	4
No Agreement	11	6	5	025	0
Total	750	367	292	84	7

Table 11: Dataset Summary

A.5 LDA Topics Correlation Matrices

Figs. 12 and 13 show correlation matrices for LDA topics.

A.6 Out-of-Domain Datasets

For moral foundation prediction, we use the dataset proposed by Johnson and Goldwasser (2018), consisting of 2K tweets by US congress members annotated for the five core moral foundations. We also use the Moral Foundation Twitter Corpus (Hoover et al., 2020), consisting of 35k tweets annotated for moral foundations. The topics across these two datasets span political issues (e.g. gun control, immigration) and events (e.g. Hurricane Sandy, Baltimore protests). Given that neither of these two datasets contain examples for the *liberty/oppression* moral foundation, we curate a small lexicon by looking for synonyms and antonyms of the words *liberty* and *oppression*. Then, we use this lexicon to annotate the congresstweets dataset [2]. We annotate a tweet as *liberty/oppression* if it contains at least four keywords, which results in around 2K tweets. The derived lex-

[2] https://github.com/alexlitel/congresstweets

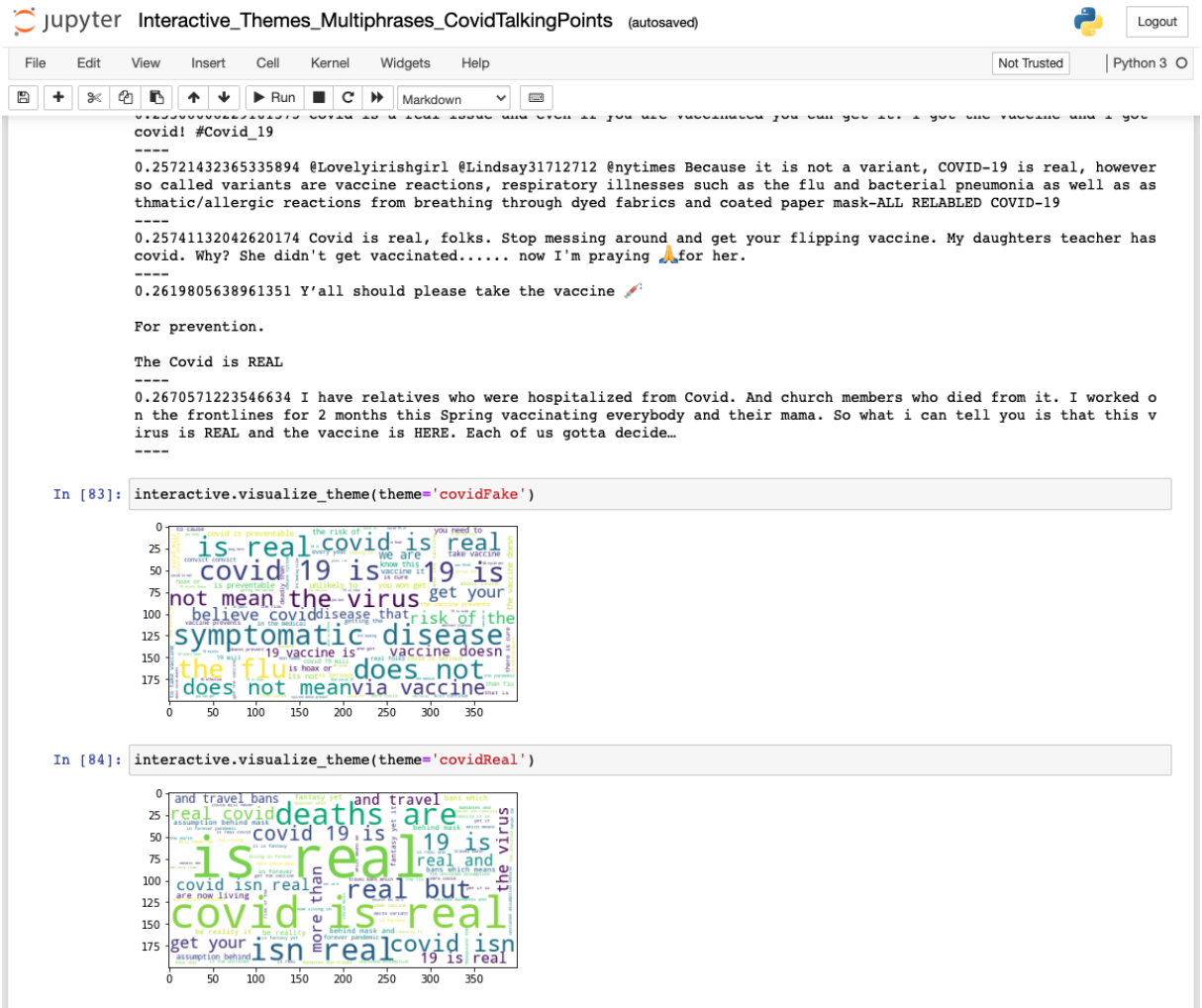
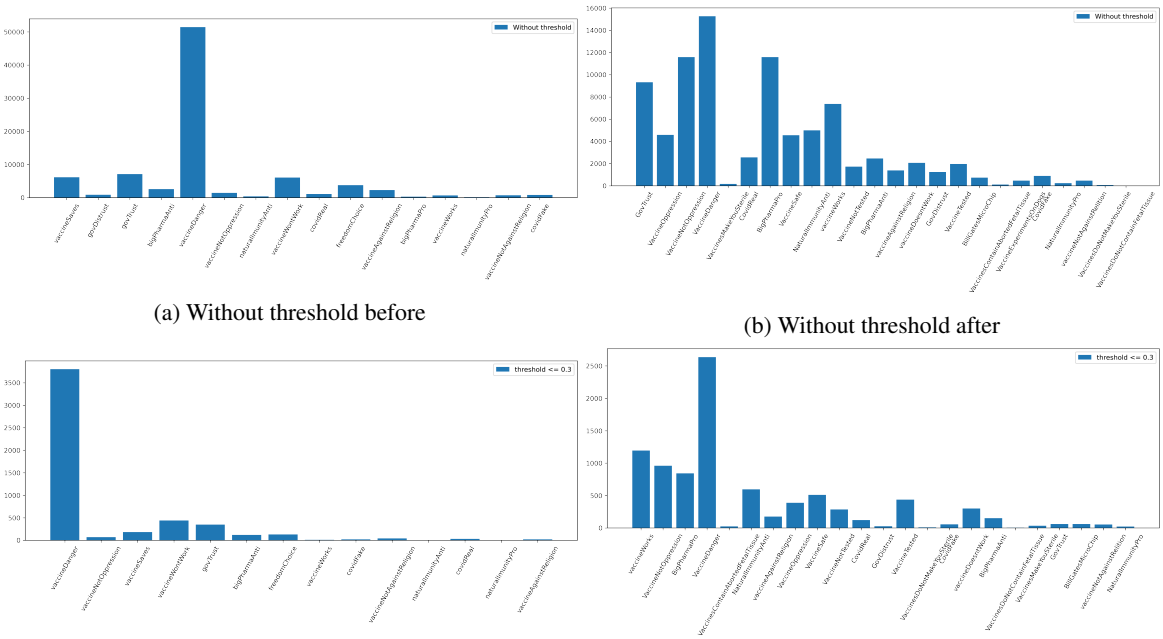


Figure 5: After querying the themes (i.e., CovidFake, CovidReal), interface shows the wordcloud.



(a) Without threshold before

(b) Without threshold after

(c) $threshold \leq 0.3$ before

(d) $threshold \leq 0.3$ after

Figure 6: Cluster assignment before and after refining arguments interactively.

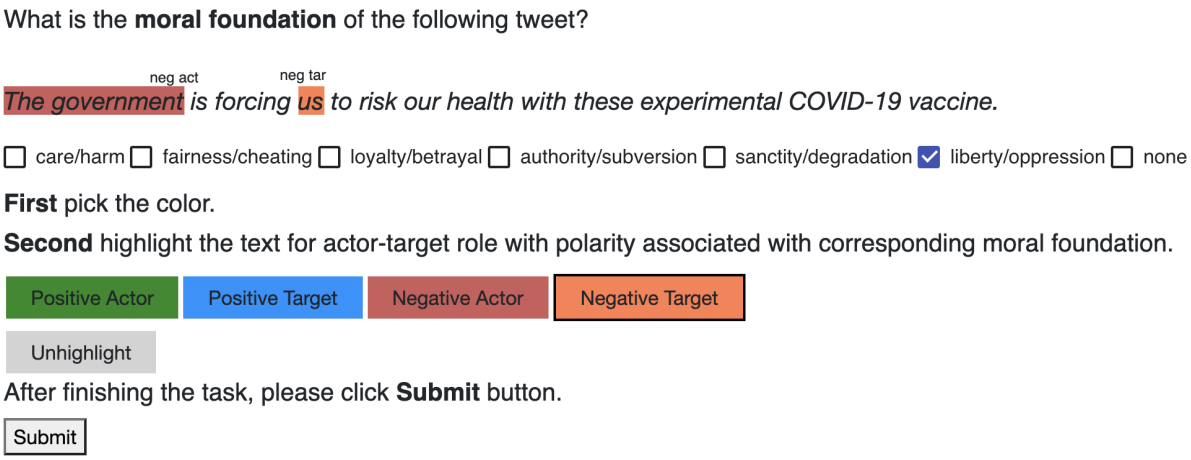


Figure 9: Annotation task interface.

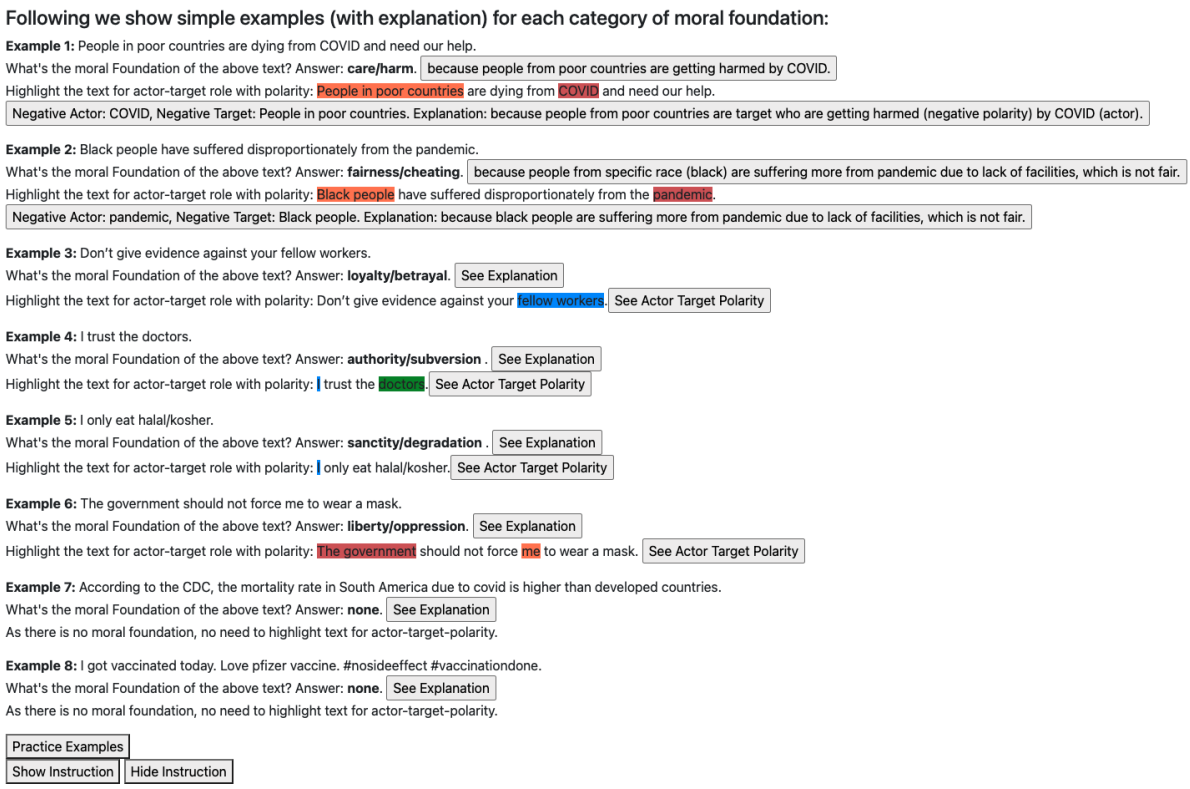


Figure 10: Examples provided to the annotators.

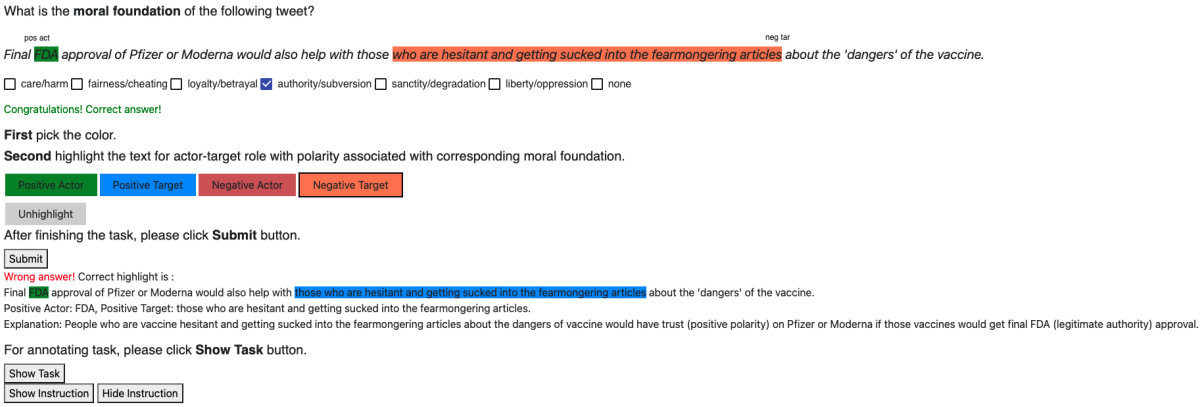


Figure 11: One of the two practice examples provided to the annotators before starting the real task.

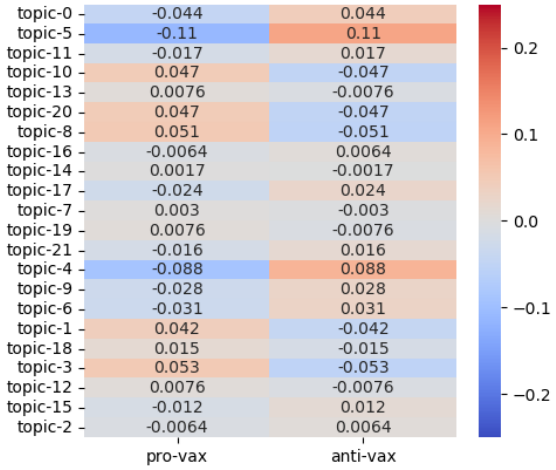


Figure 12: LDA Topics and Vax Stance

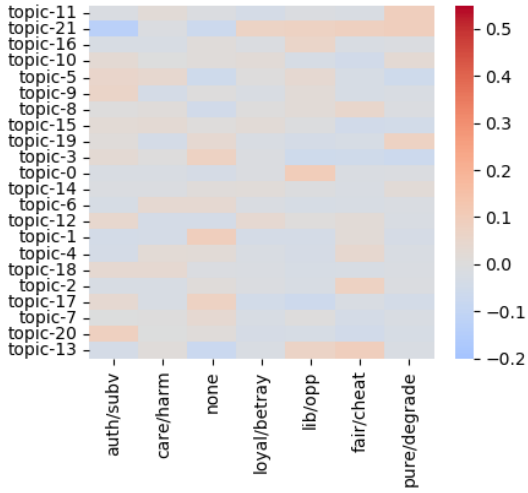


Figure 13: LDA Topics and Moral Foundations

icon for liberty/oppression can be seen in Tab. 12

To learn to predict roles, we use the subset of Johnson and Goldwasser (2018) dataset annotated for roles by Roy et al. (2021), which contains roughly 3K tweet-entity-role triplets. For polarity, we combine the Roy et al. (2021) dataset with the MPQA 3.0 entity sentiment dataset (Deng and Wiebe, 2015), which contains about 1.6K entity-sentiment pairs.

liberty, independence, freedom, autonomy, sovereignty self-government, self-rule, self-determination, home-rule civil liberties, civil rights, human rights, autarky, free-rein, latitude, option, choice, volition, democracy, oppression, persecution, abuse, maltreatment, ill treatment, dictator, dictatorship, autocracy, tyranny, despotism, repression, suppression, subjugation, enslavement, exploitation, dependence, constraint, control, totalitarianism

Table 12: Liberty/Oppression Lexicon.

For stance, we annotate our dataset of 85K unlabeled covid tweets using a set of prominent antivax and provax hashatgs. For the antivax case, we rely on the hashtags proposed by Muric et al. (2021). For the provax case, we manually annotate hashtags that have a clear provax message, and that are used in at least 50 tweets in our unlabeled dataset. The full set of hashtags used can be found in Tabs. 13 and 14.

A.7 Impact of Reasons

Tab. 15 shows the impact of explicitly modeling reasons (Eq. 5). We show the performance for the initial reasons proposed by Wawrzuta et al. (2021), which are all from the anti-vaccine perspective, and the impact of our two rounds of interaction, expanding and refining reasons (round 1) and augmenting argumentative patterns (round 2).

FullyVaccinated, GetTheVax, GetVaccinatedASAP, VaccineReady, VaxUpIL, TeamVaccine, GetTheJab, VaccinesSaveLives, RollUpYourSleeve, DontMissYourVaccine, letsgetvaccinated, TakeTheVaccine, takethevaccine, COVIDIDIOTS, SafeVaccines, ThisIsOurShotCA, LetsGetVaccinated, getthevaccine, GetVaccinated PandemicOfTheUnvaccinated, VaccineStrategy, igottheshot, vaccinationdone, ThisIsOurShot, VaccinateNiagara, TwoDoseSummer, OurVaccineOurPride, IGotMyShot, FreeVaccineForAll, VaccineEquity, COVIDIOTS, GetTheVaccine, GetVaxxed, VaccineJustice, getthejab, VaccineForAll, covidiot, gettheshot, RollUpYourSleevesMN, GoVAXMaryland, WorldImmunizationWeek, VaccinesWork, getvaccinated, GetVaccinatedNow, VaxUp, PlanYourVaccine, VaccinateEveryIndian, TakeYourShot, Vaccines4All, VaccinnateWithConfidence, firstdose, YesToCOVID19Vaccine, NYCVaccineForAll, Vaccine4All, getvaxxed, VaccinEquity,

Table 13: ProVax Hashtags

abolishbigpharma, noforcedflushots, NoForcedVaccines, ArrestBillGates, notomandatoryvaccines, betweenmeandmydoctor, NoVaccine, bigpharmafia, NoVaccineForMe, bigpharmakills, novaccinemandates, BillGatesBioTerrorist, parentalrights, billgatesevil, parentsoverpharma, BillGatesIsEvil, saynotovaccines, billgatesisnotadoctor, stopmandatoryvaccination, billgatesvaccine, cdcfraud, cdctruth, v4vglobaldemo, cdcwhistleblower vaccinationchoice, covidvaccineispoison, VaccineAgenda depopulation, vaccinedamage, DoctorsSpeakUp, vaccinefailure, educateb4uvax, vaccinefraud, exposebillgates, vaccineharm, forcedvaccines, vaccineinjuries, Fuckvaccines, vaccineinjury idonotconsent, VaccinesAreNotTheAnswer, informedconsent, vaccinesarepoison, learntherisk, vaccinescause, medicalfreedom, vaccineskill, medicalfreedomofchoice, momsofunvaccinatedchildren, mybodymychoice

Table 14: AntiVax Hashtags

MODEL	MF	VAX. STANCE
ALL (-Reasons)	60.07	67.72
+ Reasons-Original	61.51	72.62
+ Reasons-Interaction-1	61.21	**73.83**
+ Reasons-Interaction-2	**62.27**	72.53

Table 15: Contribution of reasons at different interaction rounds (Weighted F1)