A Probabilistic Framework for Analyzing Moral Perspectives in 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, it requires understanding the reasoning behind 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. Our experiments show that our framework can robust classifiers 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 implication, 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.

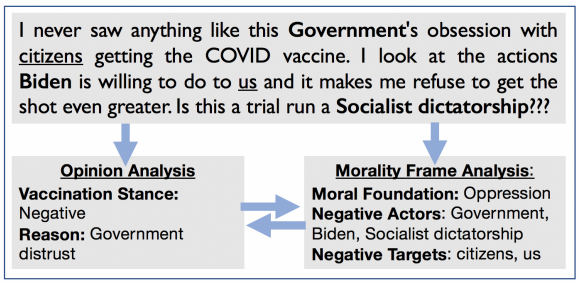


Figure 1: Holistic Analysis Framework of Social Media Posts, Connecting entity-level Moral Perspectives, Stance and Arguments Justifying it.

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. In this paper 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. 4.3

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 `"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 4.4.

3. How can text analysis models be adapted to this highly dynamic domain, without 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 their instantiation on a given debate topic and as a result might not generalize well. Instead of retraining the model from scratch, we hypothesize that given an initial model, constructed using out-of-domain data, modeling the interaction between reasons, stances and moral foundation will help enhance the initial model and provide acceptable performance. We study these settings, along with the fully supervised setting in Sec. 5.

2 Related Work

Identifying stances and arguments supporting them is a central challenge of argumentation mining (Habernal et al., 2018; Lawrence and Reed, 2020), and several works studying it in the context of the vaccine debate (Walker et al., 2014; Torsi and Morante, 2018; Morante et al., 2020), including on social media (Glandt et al., 2021). In recent years, as Covid-19 has become a central topic of discussion on social media, several works analyzed opinions and misiformation on these platform (Nguyen et al., 2020; Biester et al., 2020; Tagliabue et al., 2020; Wei et al., 2020; Kleinberg et al., 2020; Abdul-Mageed et al., 2021; Alam et al., 2021; Weinzierl et al., 2021).

Moral Foundation Theory (Haidt and Joseph, 2004; Haidt and Graham, 2007) has been widely adopted by social scientists to analyze attitudes on a wide range of topics, including political and social behaviors (Dehghani et al., 2016; Mooijman et al., 2018), as well as health and well-being choices (Pagliaro et al., 2021; Díaz and Cova, 2021; Chan, 2021). Several works studied how moral foundation theory can be operationalized in newswire and social media (Garten et al., 2016; Johnson and Goldwasser, 2018; Lin et al., 2018; Hoover et al., 2020b; Xie et al., 2019; Roy et al., 2021). Our work is also related to entity-centric affect analysis (Deng and Wiebe, 2015a; Field and Tsvetkov, 2019; Park et al., 2020; Roy et al., 2021).

Probabilistic inference using neural nets was explored in the context of traditional NLP tasks such as parsing (Chen and Manning, 2014; Weiss et al., 2015; Andor et al., 2016), named entity recogni-

tion (Lample et al., 2016) and sequence labeling systems (Ma and Hovy, 2016; Zhang et al., 2017), as well as argumentation mining (Niculae et al., 2017; Widmoser et al., 2021), and event/temporal relation extraction (Han et al., 2019). Our work is also broadly related to interactive approaches that involve humans in the training loop (Lertvittayakumjorn et al., 2020; Wang et al., 2021).

3 COVID-19 Morality Frames

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). 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 other things.

3.1 Data Collection and Annotation

There is no existing corpus of COVID-19 vaccine arguments annotated for moral foundations or morality frames, so we collected and annotated our own data set. First, we searched for tweets between April 2021 and October 2021 mentioning specific keywords, such as "covid vaccine" and "vaccine mandate". The full list of keywords can be seen in Appendix A.1, Table 6.

Then, we created an exclusive web application for annotating our unique task. Our task is to find out the moral foundation of a tweet, corresponding to one of six moral principles (e.g., "I give to the poor" expresses **care**), and then highlight the entities in the text according to (1) their roles - **actor** (a 'do-er') whose actions influence the **target** (the 'do-ee'), and (2) polarity, depending on the **positive** or **negative** influence of these actions. For example, "I give to the poor", "I" is a **positive actor**, and "the poor" is a **positive target** (benefiting from the actor's actions). On the other hand "We are suffering from pandemic" expresses **harm** as moral principles where "pandemic" is a **negative actor**, and "we" is a **negative target** (suffering from the actor's actions). We annotate our dataset using three in-house annotators pursuing Ph.D. program in Computer Science, to construct the first public COVID-19 corpus annotated with moral foundation and roles associated with the corresponding moral foundation.

3.1.1 Task Interface Details

To ensure quality work, we provide eight examples covering six moral principles and non-moral cases. The examples provided to the annotators are also provided in the Appendix A.2, Fig. 6. We provide two practice examples resembling the real task for the annotators (see the Appendix A.2, Fig. 7). Before starting the annotation task, the annotators must read the instructions, go through the examples, and practise two practice examples. Fig. 5 shows the part of our task interface. We describe the details of the annotation steps in Appendix A.2.

3.1.2 Quality Assurance

At first, we set up our task on Amazon Mechanical Turk. Next, we release multiple batches of tweets for annotation and receive poor annotation performance. Later, we decide to choose in-house annotators for our task and release a small subset of tweets for annotation. Based on the annotation quality, we select three in-house annotators. We award the annotators $ 0.75 per tweet and bonus ($2 * \$0.75 = \$1.5$) for completing two practice examples. Our work is Institutional Review Board (IRB) approved, and we follow their protocols.

Moral Foundation	Num. Tw.	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 1: Dataset Summary

TOP ANTIVAX	TOP PROVAX
(Fauci, actor, neg)	(I, actor, pos)
(People, target, neg)	(vaccine, actor, pos)
(Biden, actor, neg)	(COVID, actor, neg)
(I, target, neg)	(we, target, neg)
(they, actor, neg)	(black people, target, neg)

Table 2: Top 5 (Ent, Role, Polarity) triplets for stance

Inter-annotator agreement We calculate the agreement among annotators using Krippendorff's α (), where $\alpha = 1$ suggests perfect agreement, and $\alpha = -1$ suggests inverse 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 $\alpha = 67.15$.

3.2 Resulting Dataset

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. For moral foundation and stance, we take a simple majority vote. The final dataset statistics can be observed in Tab. 1.

To evaluate the correlation between moral foundations and stance, we calculate the Pearson correlation matrix and present it in Fig. 2. We can observe that there is a positive correlation between the anti-vax stance and the *liberty/oppression*, the *authority/subversion*, and *purity/degradation* moral foundations. In the case of the pro-vax stance, there is a positive correlation with the *care/harm* and *loyalty/betrayal* moral foundations.

In Tab. 2 we show the top five (E,R,P) tuples for each stance. We can see that the anti-vax side criticizes authority figures like Biden and Fauci, and puts self as a negatively affected entity. Meanwhile, the pro-vax side portrays the vaccine and self as good actors, and portrays minority groups as a negatively affected entity.

4 Model

In this section, we define our model to predict moral perspectives in the COVID-19 vaccine debate. We account both for supervised and weakly supervised settings. In the supervised case, we assume we have in domain training data for morality frames and stance. In the weakly supervised case, we use no direct in domain supervision.

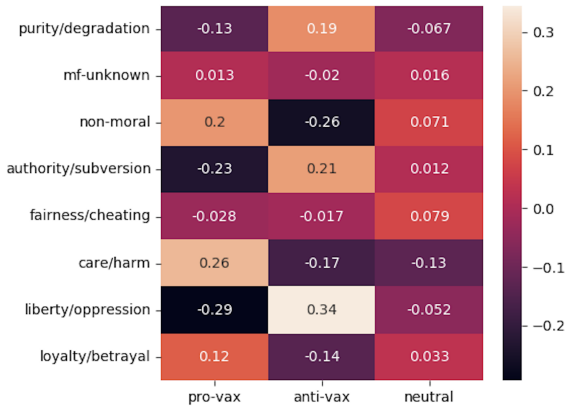


Figure 2: Pearson's corr. between MFs and Stance

4.1 Modeling Morality Frames

We define the following classifiers for predicting morality frames. Our framework is architecture-agnostic so in principle, any text classifier can be used. We specify the details of the classifiers used in this paper in Sec. 5.

Supervised Learning In the supervised case, we learn four different classifiers directly from the annotated data introduced in Sec. 3. We break down the task of predicting morality frames into four sub-tasks. For each tweet, we predict whether it is making moral judgement or not and its prominent moral foundation. For each entity in the tweet, we predict whether it is a target or a role, and whether it has positive or negative polarity.

Out-of-Domain Classifiers for Morality Frames To learn to predict morality frames in the weakly supervised case, we use out-of-domain classifiers for all tasks. 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., 2020a), 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 [1]. We annotate a tweet as *liberty/oppression* if it contains at least four keywords, which results in around 2K tweets. The derived lexicon can be observed in Appendix A.3.

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, 2015b), which contains about 1.6K entity-sentiment pairs.

4.2 Modeling Opinions

To model opinions, we define a stance classifier and a clustering method to identify repeating arguments in the COVID vaccine debate. For both methods, we rely on an unlabeled dataset of 3M tweets containing the phrase "covid vaccine" between January and October of 2021. We collected this dataset using the Twitter Academic Search API.

Stance For the supervised case, we use a classifier directly over the annotated data. For the weakly supervised case, we annotate a subset of our 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 Appendix A.4.

Common Arguments We build on the work by Wawrzuta et al. (2021), who identified common themes in the vaccination discourse online, including arguments such as "covid is not real", and "the vaccine was not properly tested". To start, we directly model the 13 themes that they suggest. To represent them, we use the textual explanation that they provide and extract its SBERT embedding (Reimers and Gurevych, 2019). We then cluster tweets based on the theme it is most similar to.

4.3 Refining Arguments Interactively

Human in the loop of arguments expressed as text. We build an interactive interface to understand COVID-19 talking points in social media. We select 24 themes with multiple phrases for analyzing our unlabeled dataset. For example, we have a theme named '**GovDistrust**' and phrase under this theme is *"lack of trust in the government"*. The full list of themes and phrases are in Appendix A.5, Table 10. We take the first phrase of 8 themes (AntiVax) from here (Wawrzuta et al., 2021), then we add multiple phrases iteratively. We expand themes related to ProVax (i.e., '**GovTrust**') and conspiracy theory (i.e., '**BillGatesMicroChip**'). We show our interactive task interface in Appendix A.5, Fig. 8 and Fig. 9.

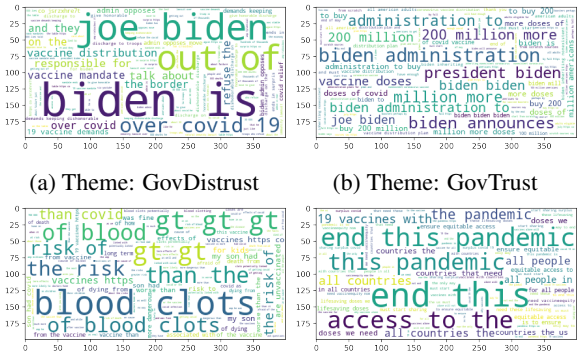
In this task, we use sentence BERT (Reimers and Gurevych, 2019) for creating embedding of unlabeled tweets and phrases. To explore tweets that are closer to the phrase embedding, we create cluster based on minimum distance (maximum similarity) and calculate cluster purity using Silhouette coefficient (Rousseeuw, 1987). Then, we assign threshold for number of assigned tweets per cluster based on closest distance ($threshold \leq [0.2, 0.3, 0.4, 0.5]$). Bar plots for cluster assignment without threshold and $threshold \leq 0.3$ both for before and after refining arguments interactively are provided in Appendix A.5, Fig. 10. To visualize talking points per theme in wordcloud, we choose top 100.

Fig. 3 shows the wordcloud of 4 themes, i.e., GovDistrust, GovTrust, VaccineDanger, and VaccineSafe having one phrase only. After adding multiple phrases, e.g., phrases with strong word for 'GovDistrust' (*"The government is a total failure "*); hedging phrases for 'GovTrust' (*"The government can be corrupt, but they are telling the truth about the covid vaccine"*), we obtain improved wordcloud (Fig. 4). We show the talking points of conspiracy theory in Appendix A.5, Fig. 11.

4.4 Joint Probabilistic Model

We propose a joint probabilistic model that reasons about morality frames, stances, the arguments made, 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 methods to model a joint distribution over relational data. This hybrid modeling paradigm allow us to leverage expressive textual encoders, and to

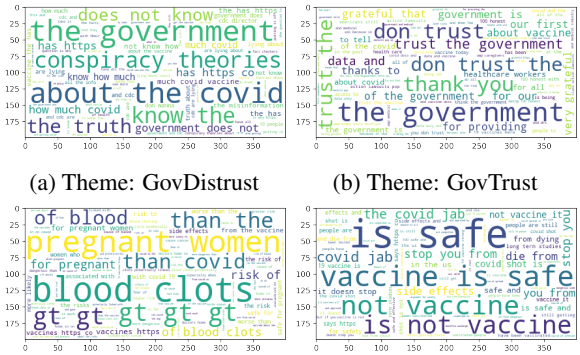
[1] https://github.com/alexlitel/congresstweets



(a) Theme: GovDistrust (b) Theme: GovTrust

(c) Theme: VaccineDanger (d) Theme: VaccineSafe

Figure 3: Wordcloud for themes and talking point before refining arguments interactively.



(a) Theme: GovDistrust (b) Theme: GovTrust

(c) Theme: VaccineDanger (d) Theme: VaccineSafe

Figure 4: Wordcloud for themes and talking point after refining arguments interactively.

introduce contextualizing information and model different interdependent decisions. Statistical relational learning methods have proven effective to model domains with limited supervision (Johnson and Goldwasser, 2018; Subramanian et al., 2018), and approaches that combine neural networks and statistical relational learning techniques 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. The explanation of how these functions are learned can be found in Section 4.5.

Base rules: 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 stance with respect to the vaccine debate.

$$r_0 : \mathtt{Tweet}(\mathtt{t_i}) \Rightarrow \mathtt{IsMoral}(\mathtt{t_i})$$
$$r_1 : \mathtt{Tweet}(\mathtt{t_i}) \Rightarrow \mathtt{HasMF}(\mathtt{t_i}, \mathtt{m})$$
$$r_2 : \mathtt{Tweet}(\mathtt{t_i}) \Rightarrow \mathtt{IsProVax}(\mathtt{t_i})$$

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).

$$r_3 : \mathtt{Tweet}(\mathtt{t_i}) \wedge \mathtt{Mentions}(\mathtt{t_i}, \mathtt{e_i}) \Rightarrow \mathtt{HasRole}(\mathtt{e_i}, \mathtt{r})$$
$$r_4 : \mathtt{Tweet}(\mathtt{t_i}) \wedge \mathtt{Mentions}(\mathtt{t_i}, \mathtt{e_i}) \Rightarrow \mathtt{HasPolarity}(\mathtt{e_i}, \mathtt{p})$$

These five base rules correspond to the classifiers introduced in Sections 4.1 and 4.2.

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 prominent moral foundation of the tweet.

$$r_5 : \mathtt{Tweet}(\mathtt{t_i}) \wedge \mathtt{Mentions}(\mathtt{t_i}, \mathtt{e_j}) \wedge \mathtt{HasRole}(\mathtt{e_i}, \mathtt{r}) \wedge \mathtt{HasPolarity}(\mathtt{e_i}, \mathtt{p}) \Rightarrow \mathtt{HasMf}(\mathtt{t_i}, \mathtt{m})$$

Dependency between stances and moral foundations: As we showed in Section 3.2, 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 moral foundation. To capture this, we model the dependency between the stance of a tweet and its moral foundation.

$$r_6 : \mathtt{Tweet}(\mathtt{t_i}) \wedge \mathtt{HasStance}(\mathtt{t_i}, \mathtt{s}) \Rightarrow \mathtt{HasMf}(\mathtt{t_i}, \mathtt{m})$$

Dependency between arguments and moral foundations/stances: Explicitly modeling the dependency between recurring arguments 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 arguments and moral foundations, and one between arguments and stances.

$$r_7 : \mathtt{Tweet}(\mathtt{t_i}) \wedge \mathtt{Mentions}(\mathtt{t_i}, \mathtt{a}) \Rightarrow \mathtt{HasMf}(\mathtt{t_i}, \mathtt{m})$$
$$r_8 : \mathtt{Tweet}(\mathtt{t_i}) \wedge \mathtt{Mentions}(\mathtt{t_i}, \mathtt{a}) \Rightarrow \mathtt{HasStance}(\mathtt{t_i}, \mathtt{s})$$

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 moral foundation should be assigned to it.

$$c_0 : \texttt{Tweet}(\texttt{t}_\texttt{i}) \wedge \texttt{IsMoral}(\texttt{t}_\texttt{i}) \Rightarrow \neg\texttt{HasMf}(\texttt{t}_\texttt{i}, \texttt{none})$$
$$c_1 : \texttt{Tweet}(\texttt{t}_\texttt{i}) \wedge \neg\texttt{IsMoral}(\texttt{t}_\texttt{i}) \Rightarrow \texttt{HasMf}(\texttt{t}_\texttt{i}, \texttt{none})$$

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 debate, we use the same rationale.

$$c_3 : \texttt{Tweet}(\texttt{t}_\texttt{i}) \wedge \texttt{Tweet}(\texttt{t}_\texttt{j}) \wedge \texttt{Mentions}(\texttt{t}_\texttt{i}, \texttt{e}_\texttt{i}) \wedge \texttt{Mentions}(\texttt{t}_\texttt{j}, \texttt{e}_\texttt{j}) \wedge \texttt{SameStance}(\texttt{t}_\texttt{i}, \texttt{t}_\texttt{j}) \wedge \texttt{HasPolarity}(\texttt{e}_\texttt{i}, \texttt{p}) \Rightarrow \texttt{HasPolarity}(\texttt{e}_\texttt{j}, \texttt{p})$$

4.5 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 $\texttt{t}_\texttt{i}$, we would like the weight of rule instance $\texttt{IsTweet}(\texttt{t}_\texttt{i}) \Rightarrow \texttt{HasMf}(\texttt{t}_\texttt{i}, \texttt{care/harm})$ to be greater than the weight of rule instance $\texttt{IsTweet}(\texttt{t}_\texttt{i}) \Rightarrow \texttt{HasMf}(\texttt{t}_\texttt{i}, \texttt{loyalty/betrayal})$. 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, and thus 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. Additional details can be found in the original paper (Pacheco and Goldwasser, 2021).

Learning in the Weakly Supervised Case To learn DRaiL models without any direct supervision, we use an Expectation-Maximization style protocol, outlined in Algorithm 1. We initialize the parameters of the neural networks for the base rules using the weakly supervised classifiers defined above, and all other rule parameters randomly. Then, we alternate between MAP inference to refine the training labels, and training the neural nets.

Algorithm 1 *Weakly Supervised Learning Protocol*

```
Random initialization for all θ_r
for r ∈ base rules do
    θ_r ← weak classifier
end for
while not converged do
    Y_gold ← MAP inference
    Train all rules locally using Y_gold
end while
```

5 Experimental Evaluation

The goal of our joint probabilistic framework is to identify morality frames and opinions in tweets by modeling them jointly. In addition to this, we want to be able to do so when there is no available direct in domain supervision. In this section, we perform an exhaustive experimental analysis to evaluate the performance of our model and each of its components.

5.1 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 of our base classifiers, both supervised and weakly supervised. For the rules that model dependencies (r_5-r_8), we concatenate the CLS token with a 1-hot representation of the symbols on the left hand side of the rule (i.e. role, sentiment, stance and argument theme), before passing it through a classifier. For rules that have the entity on the left-hand side (r_3, r_4, r_5), 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 weakly supervised models using the protocol outlined in Algorithm 1. In both cases, we used a learning rate of $2e-5$, a maximum sequence length of 100, and the AdamW optimization algorithm. In all experiments shown, we perform 5-fold cross-validation and report the micro averaged results.

5.2 General Results

Tab. 3 shows our general results for morality frames. We evaluate our standalone classifiers for both the supervised and weakly supervised case, and show the impact of modeling dependencies and

MODEL	MORAL/NM		MF		ACTOR/TARGET		POLARITY	
	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
Majority Class	37.05	43.62	8.33	23.98	34.63	36.69	46.54	58.15
Lexicon Matching			25.28	35.85	-	-	-	-
Weakly Supervised Classifiers	69.77	68.88	28.79	41.27	71.94	72.05	63.88	74.30
EM + ALL Deps. and Constr.	**78.87**	**79.71**	**36.89**	**58.86**	**83.62**	**83.83**	**76.78**	**79.71**
Supervised Classifiers	68.94	69.71	35.28	42.92	**84.71**	**84.75**	**72.92**	**84.31**
+ ALL Deps. and Constr.	**80.53**	**81.17**	**53.29**	**62.27**	84.60	84.64	71.41	83.26

Table 3: General Results (F1 Scores). MC: Morality Constraint, SPC: Stance-Polarity Constraint

MODEL	MF
ALL (-Args)	60.07
+ Args-Original	61.51
+ Args-Both-Sides	61.21
+ Arg-Interaction	**62.27**

Table 4: Contribution of Arguments to Moral Foundation Prediction for the Supervised Case (Weighted F1)

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**
+ArgMF	69.71	53.15	84.75	84.31
+ArgMF+MC	72.60	53.41	84.75	84.31
+ArgStance+SPC	69.71	42.92	84.64	83.26
+ ALL	**81.17**	**62.27**	84.64	83.26

Table 5: Ablation Study for the Supervised Case (Weighted F1). MC: Morality Constraint, SPC: Stance-Polarity Constraint

constraints using DRaiL. In both cases, modeling in a significant improvement in performance for morality and moral foundation. In the supervised case, the role and polarity numbers remain stable, while in the weakly supervised they improve considerably. By leveraging inference and our EM-style learning protocol, we are able to get a model that is fairly competitive without any in-domain direct supervision. However, note that the difference between the macro and weighted F1 scores for MFs is considerably higher for the weakly supervised case. This is because our initial out-of-domain classifier never learns to predict *loyalty/betrayal*, and we can never recover from this.

Tab. 4 shows the impact of themes and arguments ($r7 - r8$) in our model. We show the performance for the initial themes proposed by (Wawrzuta et al., 2021), which are all from the antivax perspective, the impact of expanding them with the opposing arguments, and then the impact of our interaction protocol to augment phrases. We can see that we are able to improve performance by refining arguments interactively.

5.3 Ablation Study

We show an ablation study in Tab. 5 for the supervised case. First, we can see how all dependencies contribute to the performance improvement, role to moral foundation being the most impactful. In addition to this, we can see that explicitly modeling morality constraints improves both the morality prediction and the moral foundation prediction, suggesting the advantage of breaking this decision into two modules and join them through constrained inference. We observe that the stance-polarity constraint does not have a significant impact, but does not hurt performance either, suggesting that our classifiers already captures this on their own. Lastly, we can see that the performance for roles and polarity remains stable, suggesting that these predictions support moral foundation and moral prediction, but the effect is not symmetric, potentially because the role and polarity classifiers have a very strong starting point.

6 Summary

We introduce a holistic framework for analyzing social media posts and test it on the COVID-19 vaccinate debate on Twitter. We propose a joint probabilistic framework to model morality frames and opinions, and show that we can obtain competitive performance in the supervised case. In addition to this, we show that using our framework and leveraging indirect supervision, we also obtain competitive performance when we have no direct in-domain supervision.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, El Moatez Billah Nagoudi, Dinesh Pabbi, Kunal Verma, and Rannie Lin. 2021. Mega-COV: A billion-scale dataset of 100+ languages for COVID-19. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3402–3420, Online. Association for Computational Linguistics.

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.

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2442–2452, Berlin, Germany. Association for Computational Linguistics.

Laura Biester, Katie Matton, Janarthanan Rajendran, Emily Mower Provost, and Rada Mihalcea. 2020. Quantifying the effects of COVID-19 on mental health support forums. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

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

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. Association for Computational Linguistics.

Morteza Dehghani, Kate Johnson, Joe Hoover, Eyal Sagi, Justin Garten, Niki Jitendra Parmar, Stephen Vaisey, Rumen Iliev, and Jesse Graham. 2016. Purity homophily in social networks. *Journal of Experimental Psychology: General*, 145(3):366.

Lingjia Deng and Janyce Wiebe. 2015a. Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 179–189, Lisbon, Portugal. Association for Computational Linguistics.

Lingjia Deng and Janyce Wiebe. 2015b. 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.

Anjalie Field and Yulia Tsvetkov. 2019. Entity-centric contextual affective analysis. *arXiv preprint arXiv:1906.01762*.

Justin Garten, Reihane Boghrati, Joe Hoover, Kate M Johnson, and Morteza Dehghani. 2016. Morality between the lines: Detecting moral sentiment in text. In *Proceedings of IJCAI 2016 workshop on Computational Modeling of Attitudes*.

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.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. 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 1930–1940.

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.

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

Rujun Han, Qiang Ning, and Nanyun Peng. 2019. Joint event and temporal relation extraction with shared representations and structured prediction. 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 434–444, Hong Kong, China. 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. 2020a. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, et al. 2020b. 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.

Bennett Kleinberg, Isabelle van der Vegt, and Maximilian Mozes. 2020. Measuring Emotions in the COVID-19 Real World Worry Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Piyawat Lertvittayakumjorn, Lucia Specia, and Francesca Toni. 2020. FIND: Human-in-the-Loop Debugging Deep Text Classifiers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 332–348, Online. Association for Computational Linguistics.

Ying Lin, Joe Hoover, Gwenyth Portillo-Wightman, Christina Park, Morteza Dehghani, and Heng Ji. 2018. Acquiring background knowledge to improve moral value prediction. In *2018 ieee/acm international conference on advances in social networks analysis and mining (asonam)*, pages 552–559. IEEE.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

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)*.

Marlon Mooijman, Joe Hoover, Ying Lin, Heng Ji, and Morteza Dehghani. 2018. Moralization in social networks and the emergence of violence during protests. *Nature human behaviour*, 2(6):389–396.

Roser Morante, Chantal Van Son, Isa Maks, and Piek Vossen. 2020. Annotating perspectives on vaccination. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4964–4973.

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.

Dat Quoc Nguyen, Thanh Vu, Afshin Rahimi, Mai Hoang Dao, Linh The Nguyen, and Long Doan. 2020. WNUT-2020 task 2: Identification of informative COVID-19 English tweets. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 314–318, Online. Association for Computational Linguistics.

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument mining with structured SVMs and RNNs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 985–995, Vancouver, Canada. Association for Computational Linguistics.

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.

Chan Young Park, Xinru Yan, Anjalie Field, and Yulia Tsvetkov. 2020. Multilingual contextual affective analysis of lgbt people portrayals in wikipedia. *arXiv preprint arXiv:2010.10820*.

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*. Association for Computational Linguistics.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

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.

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.

Benedetta Torsi and Roser Morante. 2018. Annotating claims in the vaccination debate. In *Proceedings of the 5th Workshop on Argument Mining*, pages 47–56.

Vern Walker, Karina Vazirova, and Cass Sanford. 2014. Annotating patterns of reasoning about medical theories of causation in vaccine cases: Toward a type system for arguments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 1–10.

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.

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.

Jerry Wei, Chengyu Huang, Soroush Vosoughi, and Jason Wei. 2020. What are people asking about COVID-19? a question classification dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

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.

David Weiss, Chris Alberti, Michael Collins, and Slav Petrov. 2015. Structured training for neural network transition-based parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 323–333, Beijing, China. Association for Computational Linguistics.

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.

Jing Yi Xie, Renato Ferreira Pinto Junior, Graeme Hirst, and Yang Xu. 2019. Text-based inference of moral sentiment change. 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 4646–4655.

Xiao Zhang, Yong Jiang, Hao Peng, Kewei Tu, and Dan Goldwasser. 2017. Semi-supervised structured prediction with neural CRF autoencoder. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1701–1711, Copenhagen, Denmark. Association for Computational Linguistics.

A Appendix

A.1 Data Collection

The keywords used to collect tweets about the COVID-19 vaccine can observed in Table 6.

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 6: List of the keywords for data collection.

A.2 Data Annotation Task

Following are the steps for completing annotation in our task interface (See Fig. 5).

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. 6) covering six moral principles and non-moral cases to our 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. 7 shows the interface of one of the two practice examples.

A.3 Liberty/Oppression Lexicon

The derived lexicon for liberty/oppression can be seen in Tab. 7

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 7: Liberty/Oppression Lexicon.

A.4 Provax and Antivax Hashtags

Tables 8 and 9 show the hashtags used to derive the stance classifier.

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 8: ProVax Hashtags

What is the **moral foundation** of the following tweet?

neg act neg tar

The government is forcing us to risk our health with these experimental COVID-19 vaccine.

☐ care/harm ☐ fairness/cheating ☐ loyalty/betrayal ☐ authority/subversion ☐ sanctity/degradation ☑ liberty/oppression ☐ none

First pick the color.

Second highlight the text for actor-target role with polarity associated with corresponding moral foundation.

Positive Actor | Positive Target | Negative Actor | Negative Target

Unhighlight

After finishing the task, please click **Submit** button.

Submit

Figure 5: Annotation task interface.

Following we show simple examples (with explanation) for each category of moral foundation:

Example 1: People in poor countries are dying from COVID and need our help.
What's the moral Foundation of the above text? Answer: **care/harm**. because people from poor countries are getting harmed by COVID.
Highlight the text for actor-target role with polarity: People in poor countries are dying from COVID and need our help.
Negative Actor: COVID, Negative Target: People in poor countries. Explanation: because people from poor countries are target who are getting harmed (negative polarity) by COVID (actor).

Example 2: Black people have suffered disproportionately from the pandemic.
What's the moral Foundation of the above text? Answer: **fairness/cheating**. because people from specific race (black) are suffering more from pandemic due to lack of facilities, which is not fair.
Highlight the text for actor-target role with polarity: Black people have suffered disproportionately from the pandemic.
Negative Actor: pandemic, Negative Target: Black people. Explanation: because black people are suffering more from pandemic due to lack of facilities, which is not fair.

Example 3: Don't give evidence against your fellow workers.
What's the moral Foundation of the above text? Answer: **loyalty/betrayal**. See Explanation
Highlight the text for actor-target role with polarity: Don't give evidence against your fellow workers. See Actor Target Polarity

Example 4: I trust the doctors.
What's the moral Foundation of the above text? Answer: **authority/subversion** . See Explanation
Highlight the text for actor-target role with polarity: I trust the doctors. See Actor Target Polarity

Example 5: I only eat halal/kosher.
What's the moral Foundation of the above text? Answer: **sanctity/degradation** . See Explanation
Highlight the text for actor-target role with polarity: I only eat halal/kosher. See Actor Target Polarity

Example 6: The government should not force me to wear a mask.
What's the moral Foundation of the above text? Answer: **liberty/oppression**. See Explanation
Highlight the text for actor-target role with polarity: The government should not force me to wear a mask. See Actor Target Polarity

Example 7: According to the CDC, the mortality rate in South America due to covid is higher than developed countries.
What's the moral Foundation of the above text? Answer: **none**. See Explanation
As there is no moral foundation, no need to highlight text for actor-target-polarity.

Example 8: I got vaccinated today. Love pfizer vaccine. #nosideeffect #vaccinationdone.
What's the moral Foundation of the above text? Answer: **none**. See Explanation
As there is no moral foundation, no need to highlight text for actor-target-polarity.

Practice Examples
Show Instruction | Hide Instruction

Figure 6: Examples provided to the annotators.

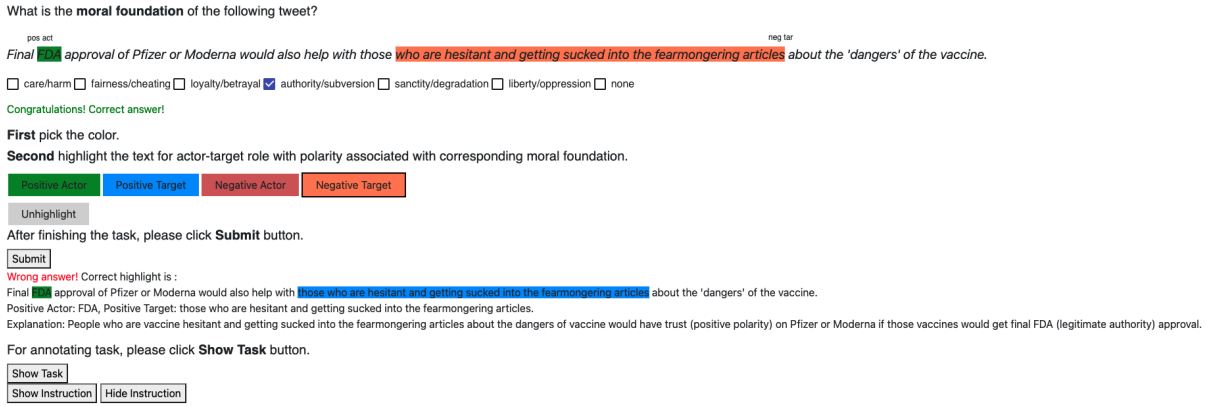
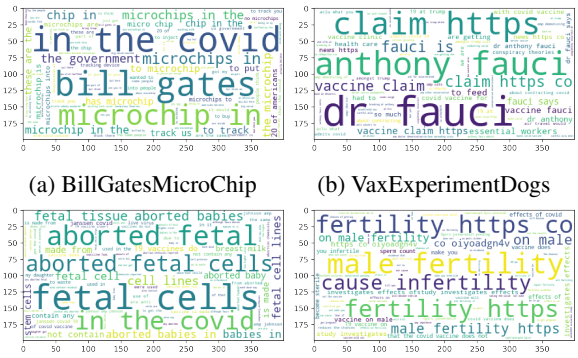


Figure 7: One of the two practice examples provided to the annotators before starting the real task.

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 9: AntiVax Hashtags



(a) BillGatesMicroChip (b) VaxExperimentDogs

(c) VaxFetalTissue (d) VaxMakeYouSterile

Figure 11: Wordcloud for themes and talking points about conspiracy theory.

A.5 Themes and Phrases

Table 10 shows the multi-phrases themes for COVID-19 talking points. We have 24 themes and each theme has multiple phrases. Interactive task interface is shown in Fig. 8 and Fig. 9.

Bar plots for cluster assignment without threshold and $threshold \leq 0.3$ both for before and after interactive session are shown in Fig. 10.

To analyze what kind of words people use in their tweets regarding conspiracy theory, we chose four common conspiracy theory themes, i.e., *'BillGatesMicroChip', 'VaxExperimentDogs', 'VaxFetalTissue', 'VaxMakeYouSterile'* and show the talking points in wordcloud (Fig. 11).

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	"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	"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"

Table 10: AntiVax Themes and phrases for COVID-19 talking points.

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"
VaxNoFetalTissue	"Vaccines were tested on fetal tissues, but do not contain fetal cells", "Vaccines do not contain aborted fetal cells"
VaxFertilityOk	"The vaccine will not make you sterile", "The covid vaccine will not affect your fertility", "No difference if fertility rate has been found between vaccinated and unvaccinated people"

Table 11: ProVax Themes and phrases for COVID-19 talking points.

jupyter Interactive_Themes_Multiphrases_CovidTalkingPoints (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

Run Markdown

Set of initial themes based on paper

In [59]:
```
#themes are the key, and phrases are the values of the dictionary

phrases = {"govDistrust": ["lack of trust in the government"],
           "vaccineDanger": ["the vaccine will be dangerous to health"],
           "covidFake": ["COVID-19 disease does not exist or is not dangerous to health"],
           "freedomChoice": ["I do not want to be vaccinated because I have freedom of choice"],
           "bigPharma": ["the vaccine was created only for the profit of pharmaceutical companies"],
           "naturalImunity": ["natural methods of protection against the disease are better than vaccines"],
           "vaccineWontWork": ["the vaccine does not exist or will not work"],
           "vaccineNotTested": ["the vaccine is not properly tested, it has been developed too quickly"],
           "vaccineExistedBefore": ["the vaccine has existed before the COVID-19 epidemic"],
           "noResponsibilityForSideEffects": ["no one is responsible for the potential side effects of the vaccine"],
           "conspiracyTheories": ["conspiracy theories, hidden vaccine effects (e.g., chips)"]
}

interactive.add_all(phrases)
```

```
Re-clustering tweets:   0%|          | 0/85799 [1:30:50<?, ?it/s]
Re-clustering tweets: 100%|██████████| 85799/85799 [00:32<00:00, 2663.37it/s]
```

Explore tweets that are closest to the theme based on distance

In [60]:
```
interactive.show_closest_tweets('freedomChoice', K=10)
```

```
0.23707894090153303 @free_rover @AhBrightWings @OccupyDemocrats It's not a choice if I'm medically exempt and my doct
ors ALL say that this vaccine is not recommended for me and many others with certain health conditions. Most people I
know who have had COVID were all vaccinated and still ended up in the hospital.I have the right to be treated
----
0.238919463419063 Quite right. We have to have choice to make our own medical decisions. If you are worried about get
ting sick take the vaccine - you have that choice too. https://t.co/GKDPqeUwlN
----
0.2486658488533775 Freedom? No control? Sure! Let's give you freedom of choice. If you CHOOSE not to get the vaccine
then you CHOOSE to accept no medical treatments for COVID symptoms! Simple
----
0.2519074511369087 I refuse the Covid vaccine because I will not have the government telling me, a woman, what to do
```

Figure 8: Interactive task interface.

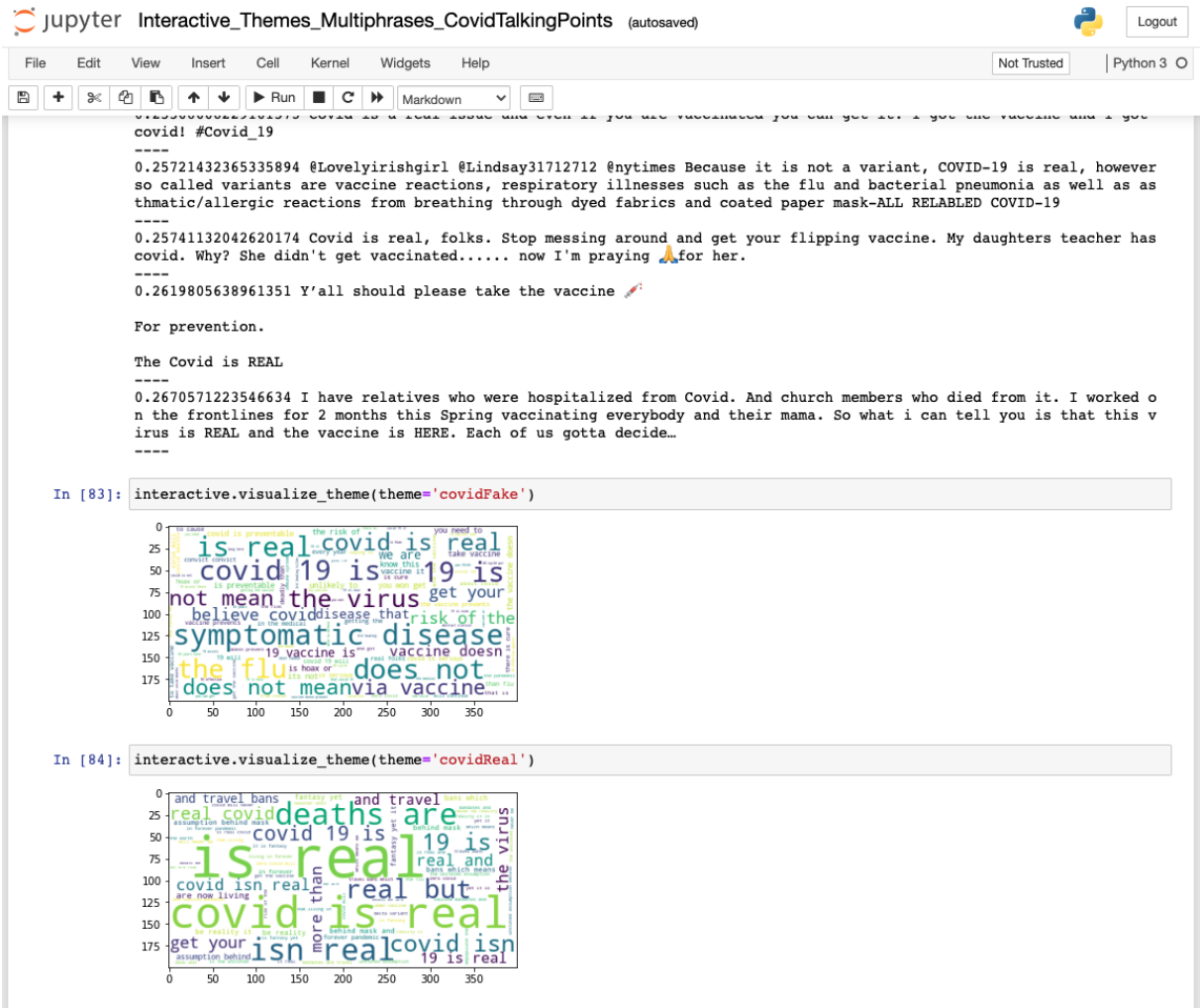
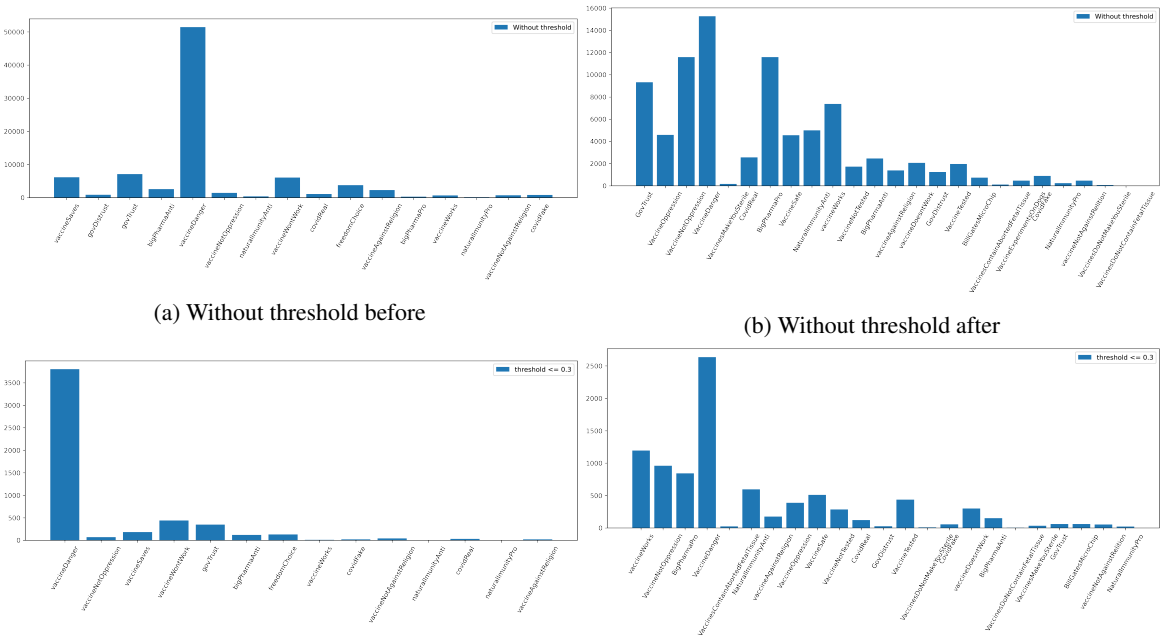


Figure 9: After querying the themes (i.e., CovidFake, CovidReal), interface shows the wordcloud.



(a) Without threshold before

(b) Without threshold after

(c) $threshold \leq 0.3$ before

(d) $threshold \leq 0.3$ after

Figure 10: Cluster assignment before and after refining arguments interactively.