Topic-Guided Stance Detection for Comparing Public Opinion Surveys with Tweets about Covid-19

Anonymous ACL submission

Abstract

Understanding public opinion, including hesitancy and scepticism towards Covid-19, is important to create appropriate public health policies. Such opinions are traditionally manually collected through surveys, though automatically measuring them through social media offers a larger reach. However, this then poses the important question of to what degree public opinion surveys and stances expressed on social media align. In this paper, we propose a new setting and method for gauging public opinion through Twitter and analysing its alignment to surveys, which we evaluate in the context of stances towards topics surrounding Covid-19 voiced by people in eight countries. Stance detection is typically framed as a pairwise sequence classification task where stance targets are provided. As this is not the case for plain tweets, we propose an alternative framing of the task, namely first identifying the tweet topic and subsequently classifying the stance towards it. To provide effective minimal supervision for training a topic-guided stance detection model, we introduce a novel topic-guided annotation technique (***TOGA***) based on unsupervised deep topic modelling and apply it to an unlabelled dataset of tweets about Covid-19. In a proxy evaluation of our method on an existing labelled stance detection dataset from the same domain (Glandt et al., 2021), we find that our few-shot method outperforms other, fully supervised approaches by 18.1 F1 points. Lastly, we show that our approach can be used effectively in conjunction with public opinion surveys for measuring public opinion and that there is a weak correlation of predicted stances with those reported in surveys.

1 Introduction

Surveys serve as an essential tool to understand public opinion on a large number of topics and are useful for creating informed public policy (Hastak et al., 2001). For instance, during the Covid-19 pandemic, the HOPE survey was conducted across countries to understand people's stances towards vaccination (Lindholt et al., 2021). However, the reach of surveys is limited – only a limited number of opinions can be taken into account.

To address these limitations, opinions expressed on social media can be leveraged. Since social media posts are open-ended, they can also provide us with relatively unadulterated insight into the topics people talk about, in contrast to surveys, where questions are explicitly drafted. Previous studies have shown that opinions expressed by the same set of people on social media and in surveys do not necessarily align (Diaz et al., 2016). Joseph et al. (2021). Compared to public surveys, opinions expressed within social media platforms tend to have stronger connotations while covering more diverse themes of public discourse. This suggests the possibility that social media captures already established and assertive opinions as opposed to public surveys, which tend to have more uncertain and hesitant responses. Hence, while stances expressed on social media cannot serve as a replacement for surveys, they can be used supplementarily, as they provide access to opinions from a much larger sample, across a wider range of topics, and at a relatively insignificant expense.

A core challenge in measuring public opinion from social media is that posts typically lack annotation of the topic of discussion, rendering existing supervised approaches (Howard and Ruder, 2018; Houlsby et al., 2019) obsolete. We thus propose a novel topic-guided stance annotation pipeline that produces weakly labeled examples, through the use of unsupervised deep topic modeling with greedy diversity sampling. Topic and stance classifiers are then trained on those examples, which are subsequently used to automatically label tweets with stances expressed on social media that we compare with the results of public opinion surveys.

For this comparison, we utilise survey response data from a study conducted by Lindholt et al.

(2021) to understand the levels and predictors of acceptance towards a government approved Covid-19 vaccine. For gauging stance towards different topics related to Covid-19, we use a large unlabelled set of 2 billion tweets (TBCOV, Imran et al. (2022)). The research questions we investigate are:

RQ1 How well can we assess public opinion from stance towards Covid-19 related topics expressed in social media?

RQ2 How do social media stances towards Covid-19 related topics vary across countries?

RQ3 Does expressed stance on social media align with public opinion surveys?

RQ4 To what extent do we observe predictors of vaccine hesitancy in social media?

In summary, our **contributions** are:

- We propose a new setting for gauging public opinions about topics from social media text through combined topic and stance prediction;
- Our proposed method for topic-guided annotation ***TOGA*** overcomes the label scarcity in unlabeled tweets and leads to an average 18.1 F1 point increase in topic and stance prediction performance, on a proxy benchmark (Glandt et al., 2021) from a similar domain;
- We provide fine-grained, semi-supervised annotations for 7 million Covid-19 related tweets across 8 countries;
- We assess the alignment between opinions expressed on social media and ones in self reported surveys across 8 countries.

2 Related Work

A variety of different approaches and task settings have been explored to perform stance detection. Stance towards a pre-defined set of topics, one at a time, is the default one. This can be done in a supervised (Mohammad et al., 2016; Augenstein et al., 2016) or an unsupervised manner (Darwish et al., 2020; Dash et al., 2022). Stance towards multiple related topics has also been explored in prior work (Sobhani et al., 2017; Allaway and McKeown, 2020). Finally, stances towards claims has been explored in Gorrell et al. (2019); Rao and Pomerleau (2022). Recently, there have been efforts to unify the different settings by combining several datasets with differing stance definitions (Schiller et al., 2021; Hardalov et al., 2021) as well as stances expressed across different languages (Hardalov et al., 2022a). An overview of these different settings of stance can be found in several surveys on the topic (Küçük and Can, 2020; ALDayel and Magdy, 2021; Hardalov et al., 2022b). Our setting differs from existing ones since we aim to identify both the topic as well as the stance from a given set of unlabelled tweets.

Close to the combined topic and stance prediction setting is work on identifying the aspects along with the designated sentiments, commonly referred to aspect-based sentiment analysis (Jang et al., 2021). The goal there is to find aspects pertaining to a particular topic along with predicting the polarities towards each aspect. Various methods have been applied within this context, ranging from deep Bi-LSTM's (Baziotis et al., 2017), Attention Networks (Yang et al., 2016; Pergola et al., 2021) to Graph Neural Networks (Zhang et al., 2019). It has also been proposed to re-frame the problem as a textual span detection task (Zhang et al., 2015; Li et al., 2018), with the aim of enriching the representations of aspects by applying a joint sequence labelling objective (Li et al., 2019) along with polarity prediction. However, in contrast to our work, most of these approaches operate in a completely supervised setting, where there is an abundance of annotated data.

3 Methods

Our overall goal is to compare stances expressed on social media about Covid-19 with those expressed in public opinion surveys. As social media data is unlabelled and no labelled stance dataset exists that covers the exact same topics as in public opinion surveys about Covid-19, going with a completely supervised setting as in prior work is impossible. Another obstacle is that prior stance detection settings (Kochkina et al., 2017; Cignarella et al., 2020) assume that topics towards which the attitude is expressed are explicitly provided. As our domain of experimentation are raw tweets (Siddiqua et al., 2019), such topic annotations do not exist.

These limitations necessitate a novel experimental pipeline. Its first component is a deep unsupervised topic model, that mitigates the lack of granular annotated data, by generating weakly supervised training sets for topic and stance classifiers (subsection 3.1). We then segment the stance detection task into a topic detection module for understanding the underlying subject within the text and a stance prediction module to designate the attitude towards the expressed topic (subsection 3.2).

3.1 Topic Classification

We follow the setting of prior work on topic classification (Lee et al., 2011; Minaee et al., 2021), framing the task as one of identifying the theme discussed within a text. This means that given a set of texts/documents $D = (d_1, \ldots d_n)$ we wish to find a set of labels $L = (l_1, \ldots l_n)$, within our topic classes $T = (t_1, \ldots t_m), l_i \in T$, for each d_i. We wish to learn a mapping $f : D \rightarrow T$ to understand the topics prevalent on social media based on their designated texts.

Recall that the overall problem setting that we are operating within does not allow for supervised training, as the raw dataset of social media texts lacks any kind of annotation. In our early experiments, we find that approaching the task in an unsupervised setting, using zero-shot prompting (Schick and Schütze, 2020a,b) or Natural Language Inference (NLI) (Wei et al., 2021) is complicated as constructing a prompt that yields adequate consistency and performance for either the topic classification or stance detection tasks is challenging (Schick et al., 2020; Liu et al., 2021).

Annotation via Topic Modeling We thus opt for using topic modeling to produce a weakly supervised set of annotations from the unlabeled set. Selecting annotated examples during task-specific finetuning is a challenging task (Shao et al., 2019), explored extensively within active learning research (Hino, 2020; Konyushkova et al., 2017). Random sampling can lead to poor generalization and knowledge transfer within the novel problem domain (Das et al., 2021; Perez et al., 2021). To mitigate the inconsistency that can be caused by choosing suboptimal examples, we propose to use deep unsupervised topic models, which allow us to sample relevant examples for each class of interest. We further enhance the model with a greedy selection process for diversity sampling (Shao et al., 2019; Yang et al., 2015) within the relevant examples generated by the topic model. The diversity maximisation sampling is modeled similar to Yang et al. (2015). We call this few-shot topic-guided annotation method *TOGA*.

The topic model we train is based on the technique proposed by Angelov (2020) that tries to find topic vectors while jointly learning document and word semantic embeddings. It is shown that learning unsupervised topics in this fashion maximizes the total information gained, about all texts D when described by all words W.

$$I(D, W) = \sum_{d \in D} \sum_{w \in W} P(d, w) \log \left(\frac{P(d, w)}{P(d)P(w)} \right)$$

This feature is very useful for finding relevant samples across varying classes, allowing us to conduct a heuristic search within the learned documents d_i, by assigning each topic class $t_i \in T$ a relevant set of keywords $(k_1 \ldots k_{l_i})$, with l_i designating the maximum amount of keywords per that class. We choose to use the verbalizers found in our early zero-shot experimentation as the keywords during this heuristic search. The keyword search yields relevancy scores $(r_1, \ldots r_n)$ for each of the documents used for training. We further refine this dataset, by searching for increasingly more diverse samples after each annotation. The search within the relevant examples is organized as follows: (1) Iteratively add the most relevant 10% of the documents per class, w.r.t their relevancy scores r_i into a set A; (2) iteratively adjust the relevancy scores r_i after each annotation, by finding the sentence that is least similar to the current set of annotated examples; (3) annotate the most relevant example w.r.t the adjusted r_i adding to the annotated set A.

To find diverse samples, in each iteration i, we find a vector v_i by averaging the representations of the annotated documents A produced by a GPT-2 model and compute a cosine similarity between v_i and the vectors representations u_j of all unannotated sequences. We adjust the relevancy score for each document according to the similarity score.

$$v_i = \frac{1}{|A|} \sum_{a \in A} PLM(a) \quad (1)$$

$$r_j = cos(\boldsymbol{v}_i, \boldsymbol{u}_j) = \frac{\boldsymbol{v}_i \cdot \boldsymbol{u}_j}{||\boldsymbol{v}_i|| \cdot ||\boldsymbol{u}_j||} \; \forall j \notin A \quad (2)$$

Next, we annotate the new most relevant sequence, adding it to A and continue this iterative annotation process to obtain at least 64 examples per class.

Few-shot setup Having generated a dataset for topic classification, we leverage the robustness of Transformer-based PLMs and finetune using the early annotated examples, casting the task into a few-shot setting. This effectively allows us to transfer the knowledge embedded within the PLM onto our problem domain. We use the fine-tuning approach by Mosbach et al. (2020); Liu et al. (2019)

to avoid the instability that can be caused by catastrophic forgetting, small-sized fine-tuning dataset or optimization difficulties.

3.2 Topic-Guided Stance Detection

Given the topic t_i for each document d_i, obtained using *TOGA*, we classify the stance expressed within that text towards the topic. We opt for a three-way stance classification setting, $S = \{\text{FAVORS, REJECTS, NEUTRAL}\}$, this being the predominant stance formulation (Rajendran et al., 2018; Glandt et al., 2021; ALDayel and Magdy, 2021). Stance detection can be generalized as pairwise sequence classification, where we learn a mapping $f : (d_i, t_i) \rightarrow S$. To learn this mapping we combine the textual sequences with the stance labels. The combination is implemented using a simple prompt commonly used for NLI tasks (Lan et al., 2019; Raffel et al., 2020; Hambardzumyan et al., 2021), where the textual sequence becomes the premise and the topic the hypothesis.

[CLS] *premise* [EOS] Stance towards *topic* [EOS]

The results of this process is a supervised dataset for stance prediction $D_{stance} = ((Prompt(d_1, t_1), s_1) \ldots (Prompt(d_n, t_n), s_n))$ where $\forall s_i \in S$. We use the topics obtained from the topic model and fine-tune a set of PLMs (see Appendix C) using Mosbach et al. (2020), to obtain the final stance detection model.

4 Experimental Setup

4.1 Data

We use four datasets in our experiments. We analyse attitudes expressed in social media data using *unlabelled Covid-19 tweets*; to validate our guided annotation technique we use a *proxy bechmark* within the same problem domain; we create an *dataset with expert annotation* for the unlabelled Covid-19 tweets; and the data from the *HOPE survey* creates the foundation for our comparison with a Covid-19 related survey.

Unlabelled Covid-19 tweets Imran et al. (2022) provide a set of $2B$ tweet IDs and metadata. The study proposes a geotagging method for obtaining geolocation information from the tweets, enabling for per country analysis. We sampled $7M$ English tweet IDs authored by users from the 8 countries mentioned above and *hydrated* them to obtain their tweet texts. The tweets are distributed as follows: Denmark – 588,127, France – 537,121, Germany – 609,968, Hungary – 9,802, Italy – 298,730, Sweden – 123,252, USA – 2,041,295, UK – 2,002,070. This results in a dataset of social media attitudes, which we further use in our prediction pipelines for obtaining the stances expressed towards topics of interest mentioned in the *HOPE survey*.

Proxy benchmark We use a dataset introduced by Glandt et al. (2021) to benchmark our annotation and prediction techniques. The dataset includes $7,122$ tweets annotated using Amazon Mechanical Turk to obtain topics and stances. The topics chosen concern attitudes regarding Anthony S. Fauci, M.D., Keeping Schools Closed, Stay at Home Orders and Wearing a Face Mask.

Expert-labelled evaluation set As the topics in the Glandt et al. (2021) dataset do not match those from the *HOPE survey*, we additionally create an expert-labeled evaluation set as follows: (1) sample a representative set of 1 million tweets randomly stratified by countries; (2) train a topic model on the sampled set; (3) use the topic model to sort the examples into high, medium and low confidence percentile buckets w.r.t the keywords provided per class, similar to the process used for ***TOGA***; (4) sample 3 examples from each bucket per class; (5) randomly shuffle the instances; (6) ask expert annotators to label the dataset.

We use two different pairs of expert annotators per each half of the annotation process. Annotators are asked to label a tweet with up to three topics and the stance towards each topic. We analyze inter-annotator agreement with two metrics: exact match, i.e. it counts as an agreement between annotators only when the first choice of both authors coincides, and soft match, if there is at least one coinciding class between the annotators for a single example, regardless of the order. For the exact match, Krippendorff's α for topics is 0.565 and for stance is 0.822. For the soft match, Krippendorff's α for topics is 0.730 and 0.683 for stance. For more fine-grained results see Table 5 in section 6. Disagreements between the annotators are resolved by discussing and merging each disagreement case creating the final evaluation set of 160 examples.

HOPE Survey The HOPE survey[1] collects $18,231$ individual survey responses from eight countries towards self-reported vaccine acceptance and other correlated factors to understanding the

[1] https://hope-project.dk/

cause for vaccine hesitancy across the different countries. The data is collected through online surveys between September 2020 and February 2021. We disregard all questions related to demographics for the purpose of our comparison. The study correlating the different factors analysed in the survey predicts major difficulties convincing vaccine sceptics, as their views often align towards overall antisystemic attitudes (Lindholt et al., 2021).

4.2 Models

We explore several PLM Transformer architectures, fine-tuning *roberta-base, roberta-large, xlm-roberta-base, xlm-roberta-large* architectures (Liu et al., 2019; Conneau et al., 2019), with a grid search along the batch sizes of $B = [8, 16, 32]$, the few-shot sizes of $[8, 16, 32, 64]$. To ensure stable models, we follow the fine-tuning procedure by (Mosbach et al., 2020), adding a linear warmup on the initial 10% of the iteration raising the learning rate to $2e-5$ and decreasing it to 0 afterwards. We use a weight decay of $\lambda = 0.01$ and train for 3 epochs with global gradient clipping on both topic classification and stance detection tasks. We find that learning for longer epochs does not yield improvement over the initial finetuning. The optimizer used for experimentation is an AdamW (Loshchilov and Hutter, 2017) with a bias correction component added for stability of the experimentation (Mosbach et al., 2020).

Topic Guidance Recall that we introduce the few-shot topic-guided annotation method ***TOGA***, which allows us to pick relevant samples per class for further fine-tuning. We evaluate its effectiveness by fine-tuning PLMs on the examples it generates and compare it with training on a random stratified sample of the same size. To further signify the importance of relevant sample selection we also perform *linear probing*, i.e. training a final classification head with a frozen PLM and comparing the results obtained with and without *TOGA*.

Model Variants We evaluate several model fine-tuning variations with and without the application of **TOGA**. Within our experiments we refer to the following models: (1) **PLM random_sample=**k - a pretrained language model that was finetuned using k random samples per class. These are used as baselines for comparisons with **TOGA**; (2) **PLM TOGA=**k - a pretrained language model finetuned on k TOGA examples per class.

We also conduct experimentation on frozen PLMs, while only training a classification head, which we designate by adding the *lin_prob* suffix.

4.3 Evaluation Metrics

To evaluate our models and have a fair comparison with the introduced benchmarks we use a standard set of metrics for classification tasks such as F1, precision, recall and accuracy.

5 Results and Analysis

We evaluate our proposed method in three settings: a proxy evaluation on an existing stance benchmark dataset (subsection 5.1), an evaluation on the expert-labeled evaluation set (subsection 5.3), and a comparison of our results to those from the HOPE survey (subsection 5.5).

5.1 Proxy benchmark assessment

Having obtained the best model and annotation configuration in the experiments described above, we compare our results with a proxy benchmark from (Glandt et al., 2021), a stance detection dataset annotated towards Covid-19 tweets, though covering different topics than those from the HOPE Survey (subsection 4.1). We use **TOGA** to sample a few-shot dataset of 64 examples per class in the benchmark, while preserving their stance labels. Note that this is $10x$ smaller than the number of examples used for training in Glandt et al. (2021). This allows us to validate the effectiveness of our overall resulting method for the specific task of automated topic and stance annotation for tweets.

As can be seen in Table 1 we are able to outperform other stance detection approaches used by Glandt et al. (2021) with an order of magnitude fewer training examples, by an average of 18.1 F1 points. For a granular overview of the experiments, see Table 4 in Appendix A.

5.2 Topic Guided Annotation and Classification

To evaluate the effect of *TOGA*, we fine-tune the few-shot classification models following section 3.1, with and without *TOGA*. This means that any experiment that is marked as *Random* used randomly sampled stratified examples. We show the effect of using *TOGA*, with a frozen PLM (linear probing) and a standard fine-tuning setup (see also subsection 4.2). In both cases our method produces competitive results, improving on the benchmark

	Ours	BERT	BERT-NS	BERT-DAN
Avg F1	**0.986**	0.810	0.818	0.815
Acc	**0.972**	0.794	0.797	0.790

Table 1: Evaluating the methods on the stance detection task from the proxy benchmark (Glandt et al., 2021)

proposed for the proxy dataset (Glandt et al., 2021) presented in Table 6 and Table 7 in Appendix A.

Few-shot fine-tuning We evaluate the effectiveness of the method in a standard few-shot setup, where we fine-tune the parameters across the whole PLM with a variety of hyperparameter configurations mentioned in Appendix A. We observe an improvement of an average of 12 points across all metrics, example amounts and architectures across 10 runs. We can therefore conclude that *TOGA* is highly effective for topic annotation and few-shot training. From these comprehensive results we choose the best training and annotation configuration for annotating the unlabelled tweets. The final topic and stance detection models are a complete fine-tune of *roberta-base* on 64 examples generated by **TOGA** per class. This model is referred to as *Our method* in further experiments.

5.3 Expert annotation benchmark

We further test our method on the expert annotated evaluation set (see section 4.1), a sample of 160 tweets from the unlabelled set. Although the amount of examples varies per class, we are still able to get a general grasp of the predictive performance on the targets of interest in Table 2. A prediction is considered correct if it exactly matches with one of the $(topic, stance)$ pairs present within the annotation set for the respective tweet.

For the subsequent analysis in subsection 5.4, we omit classes that do not have adequate representation within this benchmark, by dropping anything below the median support amount from the original set. Also, only the classes where the model achieves above 60 F1 score are considered for further analysis to ensure an empirically sound analysis, leaving 9 topics.

5.4 Social Media Stance Towards Covid-19 Across Countries

Next, we want to understand how stances towards the different Covid-19 related topics vary across countries (RQ2). To this end, we automatically label all tweets using our best method, split them

Topic	Prec.	Recall	F1	#
Trust in the NHA	0.13	1.0	0.22	8
Trust in scientists	0.75	1.0	0.86	18
Trust in government	0.65	1.0	0.79	35
Democratic rights	0.00	0.0	0.00	6
Support of protests	1.00	1.0	1.00	4
Conspiracy beliefs	0.67	1.0	0.80	10
Misinformation	1.00	1.0	1.00	11
Fatigue	0.40	1.0	0.57	6
Behaviour change	0.08	1.0	0.15	5
Knowledge	0.25	1.0	0.40	5
Concern, family	0.27	1.0	0.43	5
Concern, hospitals	1.00	1.0	1.00	10
Concern, society	0.11	1.0	0.20	12
Concern, crime	0.20	1.0	0.33	4
Concern, the economy	0.60	1.0	0.75	16
Support for restrictions	0.75	1.0	0.86	17
Vaccine Hesitancy	0.82	1.0	0.90	9

Table 2: Performance of the stance detection model, per topic on the expert annotated data-set.

by country and compare them by topic in Figure 1. While there are clear agreement across countries across the tweets (e.g., for *trust in scientists*), there are topic that show a higher divergence, such as *support of restrictions* and *vaccine hesitancy*.

5.5 Comparing Public Opinion Surveys with Social Media Data

Recall that we want to understand how opinions are expressed on Twitter, with regards to vaccine and other Covid-19 related topics. We base the topics for our analysis on the HOPE survey (Lindholt et al., 2021). RQ3 poses the question of how the stances expressed in the dataset of tweets relates to this original study. We show that there is no correlation between the social media stance of English speakers and the original survey results by country, see Table 3. As the number of data points to correlate is very small (the survey compared only eight countries) we performed the same analysis on the state level.[2] Specifically, we for each tweet extracted the address that appears in the user-description field of the tweet's author, and used a geo-location tagging tool [3] to estimate the state of the user. The survey data contained an "exact address", from which we extracted the same information. By breaking the data down to this level, we were able to calculate correlation over 95 data-points, an increase of an order of magnitude. The result of this more granular analysis again demonstrates the lack of correlation between

[2]For countries which are not divided into states (e.g., Denmark) we performed the analysis on the county or region level.

[3]https://nominatim.org/

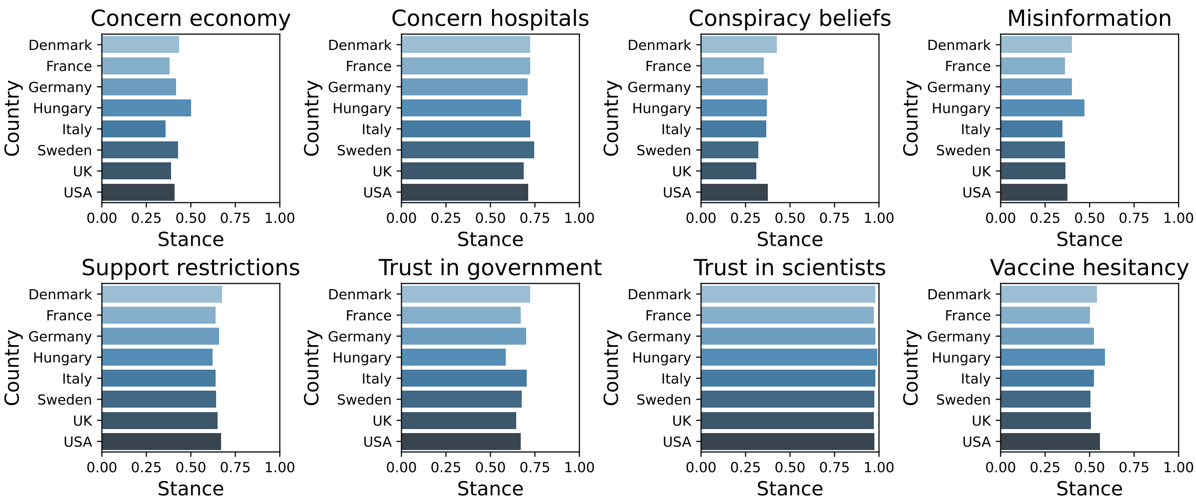


Figure 1: Comparison of the aggregated stance towards predictors for each country. “Favour” equals 1 and “Against” equals 0.

Topic	Correlation	
	Country	State
Concern, the economy	0.047	0.089
Concern, hospitals	0.071	0.073
Conspiracy beliefs	0.261	**0.645**
Misinformation	0.642	**0.319**
Support in restrictions	0.023	-0.128
Trust in the government	-0.190	**-0.364**
Trust in scientists	0.523	0.183
Vaccine hesitancy	0.047	**-0.294**

Table 3: Correlations of the Twitter stances with the survey, across countries and states. Items in bold are statistically significant (p-value < 0.05).

the two mediums and populations, with the exception of the semantically similar topics “conspiracy beliefs" and “misinformation". We leave the analysis of this phenomenon for future work.

The gap between survey results and expression of stance on social media has been previously demonstrated by Joseph et al. (2021). This discrepancy we also observe makes the addition of social media data to surveys as a data source even more important to understand overall public opinion towards a topic.

5.6 Predictors of vaccine hesitancy

The HOPE survey aims to understand which predictors influence vaccine hesitancy across cultures for individuals who participate in their survey, and we want to extend these insights to the social media data collected (RQ4). The authors of the survey calculated the correlation of the vaccine hesitancy level of the participants with the other variables that the survey had probed for. Following this, we perform an analysis of predictors of vaccine hesitancy using the collected Twitter data by correlating the aggregated level of vaccine hesitancy expressed in the tweet data with the remaining variables. We perform this analysis using three levels of granularity: the country (Figure 3a) and state (Figure 3b) levels, as in the previous section, and the individual *user* level (Figure 3c). To calculate the correlation at the user level, we first for each user collect the tweets that they authored, then split them by the main topic that our model predicted for them. Then, for each topic and for each user we calculate the aggregated stance of the user towards the topic by simple mean averaging.[4] Not every user expressed an opinion about each one of the topics. Therefore, when we correlated two topics we considered only users that tweeted about both.

As can be seen in Figure 3, each level of granularity produces a slightly different correlation profile, where the country level profile stands out as the most distinct. We attribute this to the fact that the small number of data points at the country level can introduce a high level of noise.

When comparing Figure 3c to the survey results,[5,6] the differences between stances expressed in social media and survey results becomes apparent again. Indeed, while some of the most predictive variables according to the HOPE survey are *Trust in scientists* and *Conspiracy believes*, their

[4] Here, “Favour” equals 1 and “Against” equals 0.

[5] This granularity level is the one that is most compatible with how the survey has been conducted

[6] Figure 2 in Lindholt et al. (2021)

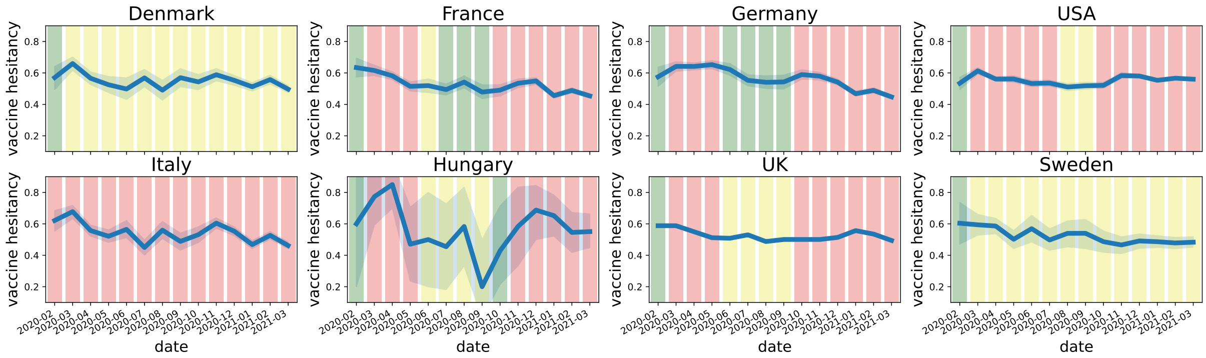
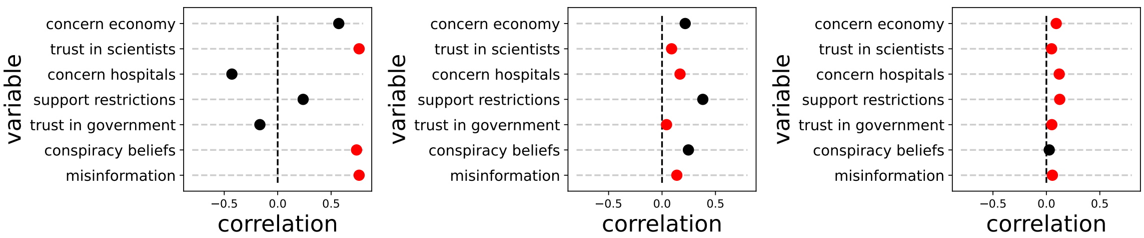


Figure 2: Development in vaccine hesitancy over time across countries. The background colour corresponds to the severity of lockdown restrictions. Green = no restrictions. Yellow = staying at home recommended. Red = lockdown in place. See Appendix E for additional restriction types.



(a) Country level

(b) State level

(c) User level

Figure 3: Predictors of vaccine hesitancy. Red markers indicate p-value < 0.05.

correlation with *Vaccine hesitancy* is almost zero according to the tweets.

Temporal analysis One of the advantages of observing stance on social media compared to surveys is that our analysis is not time-constrained and can be extended at any time by collecting new data. Therefore, we present in Figure 2 the temporal development in vaccine hesitancy by country across the whole time span of the Twitter dataset. To generate this figure, we average the stance expressed towards vaccine hesitancy for each country in each month using the tweet's timestamp field. The background colour corresponds to the severity of Covid restrictions related to face masks.[7]

Clear differences can be seen across the different countries. While some countries such as France and Germany display a steady decline in vaccine hesitancy, the trends differ strongly compared to other countries. There are no clear connections between restrictions and vaccine hesitancy, which confirms the results in Figure 3 in which we can see only a weak correlation between the *support of restrictions* and *vaccine hesitancy*. Nevertheless, these results present a starting point to further understand public opinion on Covid-19 related topics and the connection to vaccine hesitancy and global events.

[7]Taken from https://ourworldindata.org/policy-responses-covid

6 Conclusions

In this study, we propose a scalable method for gauging public opinion from social media text and assess its alignment to public opinion surveys across 8 countries. We outline an automated pipeline for semi-supervised topic and stance annotation of a large number of tweets regarding Covid-19. We find that while we can reliably assess stances towards different Covid-19 related topics from Twitter, these do not align with opinions expressed by people in online surveys. While our method does not replace surveys as a tool for measurement of public opinion, it can complement it and offer advantages like accessibility, diversification and overcoming response bias. Further, our pipeline allows for a granular analysis of the reasoning of people's stances as well as flexibility around the temporal analysis.

Limitations

At the current state, we observe the stance of English speakers across different topics. As we include countries where the main language is other than English, future work should focus on extending this study to a multilingual setup including the use of multilingual models. We think our insights are nevertheless valuable, as we can show that our approach can analyse and compare communities of a country, such as the English speaking population, and as English is a widely spoken language across all the countries included.

Further, a larger expert annotated benchmark would allow for better performance evaluation of the annotation models, consequently allowing for the discussion of a wider range of topics of interest. This improvement would propel the method for more fine-grained analysis, with consistent and robust annotation modules. Future work should address this limitation by crowd-sourcing the annotations.

References

Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Gor Arakelyan, Gevorg Soghomonyan, and The Aim team. 2020. Aim.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 747–754.

Alessandra Teresa Cignarella, Mirko Lai, Cristina Bosco, Viviana Patti, Rosso Paolo, et al. 2020. Sardistance@ evalita2020: Overview of the task on stance detection in italian tweets. *EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, pages 1–10.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. 2020. Unsupervised user stance detection on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):141–152.

Rajshekhar Das, Yu-Xiong Wang, and José MF Moura. 2021. On the importance of distractors for few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9030–9040.

Saloni Dash, Dibyendu Mishra, Gazal Shekhawat, and Joyojeet Pal. 2022. Divided we rule: Influencer polarization on twitter during political crises in india. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):135–146.

Fernando Diaz, Michael Gamon, Jake M. Hofman, Emre Kıcıman, and David Rothschild. 2016. Online and Social Media Data As an Imperfect Continuous Panel Survey. *PLOS ONE*, 11(1):1–21. Publisher: Public Library of Science.

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.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. Warp: Word-level adversarial reprogramming. *arXiv preprint arXiv:2101.00121*.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9011–9028, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022a. Few-shot cross-lingual stance detection with sentiment-based pre-training. *In Proceedings of the 36th AAAI Conference on Artificial Intelligence*.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022b. A survey on stance detection for mis- and disinformation identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277, Seattle, United States. Association for Computational Linguistics.

Manoj Hastak, Michael B Mazis, and Louis A Morris. 2001. The role of consumer surveys in public policy decision making. *Journal of Public Policy & Marketing*, 20(2):170–185.

Hideitsu Hino. 2020. Active learning: Problem settings and recent developments. *arXiv preprint arXiv:2012.04225*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Muhammad Imran, Umair Qazi, and Ferda Ofli. 2022. Tbcov: two billion multilingual covid-19 tweets with sentiment, entity, geo, and gender labels. *Data*, 7(1):8.

Hyeju Jang, Emily Rempel, David Roth, Giuseppe Carenini, Naveed Zafar Janjua, et al. 2021. Tracking covid-19 discourse on twitter in north america: Infodemiology study using topic modeling and aspect-based sentiment analysis. *Journal of medical Internet research*, 23(2):e25431.

Kenneth Joseph, Sarah Shugars, Ryan J. Gallagher, Jon Green, Alexi Quintana Mathé, Zijian An, and David Lazer. 2021. (mis)alignment between stance expressed in social media data and public opinion surveys. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 312–324. Association for Computational Linguistics.

Elena Kochkina, Maria Liakata, and Isabelle Augenstein. 2017. Turing at semeval-2017 task 8: Sequential approach to rumour stance classification with branch-lstm. *arXiv preprint arXiv:1704.07221*.

Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. 2017. Learning active learning from data. *Advances in neural information processing systems*, 30.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary. 2011. Twitter trending topic classification. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 251–258. IEEE.

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. Transformation networks for target-oriented sentiment classification. *arXiv preprint arXiv:1805.01086*.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting bert for end-to-end aspect-based sentiment analysis. *arXiv preprint arXiv:1910.00883*.

Marie Fly Lindholt, Frederik Jorgensen, Alexander Bor, and Michael Bang Petersen. 2021. Public acceptance of covid-19 vaccines: cross-national evidence on levels and individual-level predictors using observational data. *BMJ open*, 11(6):e048172.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: a comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3):1–40.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *Advances in Neural Information Processing Systems*, 34:11054–11070.

Gabriele Pergola, Elena Kochkina, Lin Gui, Maria Liakata, and Yulan He. 2021. Boosting low-resource biomedical qa via entity-aware masking strategies. *arXiv preprint arXiv:2102.08366*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Gayathri Rajendran, Bhadrachalam Chitturi, and Prabaharan Poornachandran. 2018. Stance-in-depth deep neural approach to stance classification. *Procedia computer science*, 132:1646–1653.

Delip Rao and Dean Pomerleau. 2022. Fake News Challenge.

Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically identifying words that can serve as labels for few-shot text classification. *arXiv preprint arXiv:2010.13641*.

Timo Schick and Hinrich Schütze. 2020a. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Timo Schick and Hinrich Schütze. 2020b. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance Detection Benchmark: How Robust is Your Stance Detection? *KI - Künstliche Intelligenz*, 35(3):329–341.

Jingyu Shao, Qing Wang, and Fangbing Liu. 2019. Learning to sample: an active learning framework. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 538–547. IEEE.

Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono. 2019. Tweet stance detection using an attention based neural ensemble model. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*, pages 1868–1873.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557, Valencia, Spain. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G Hauptmann. 2015. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113(2):113–127.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks. *arXiv preprint arXiv:1909.03477*.

Meishan Zhang, Yue Zhang, and Duy Tin Vo. 2015. Neural networks for open domain targeted sentiment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 612–621.

Appendix

A TOGA with few shot experimentation

To evaluate our method we propose a varying set of experiments, that includes experimentation using a frozen and unfrozen PLM along with a $k = [4, 8, 16, 32, 64]$ examples per target class, that were generated either using ***TOGA*** or randomly sampled in a stratified manner. We complete a grid search along these configurations presented in Table 7 and Table 4, evaluating on the dataset from (Glandt et al., 2021). This allows us to gauge both an in-depth overall assessment of the method performance, along with a granular understanding about model generalisation and robustness towards the designated classes. All of the experimentation is tracked using Aim (Arakelyan et al., 2020), which we use to obtain the optimal configuration for training and annotation.

Linear probing In this set of experiments, we freeze the parameters in the PLM and fine-tune only using the new classification head. This evaluation method allows us to gauge the immediate effect that the training set created with ***TOGA*** has on the final results found in Table 7. It is apparent that regardless of the chosen architecture and the number of examples per class provided during the fine-tuning process, the results obtained by training on the examples provided by ***TOGA*** are vastly superior compared to training on random stratified samples. We are able to obtain an increase of $5F1$ points, averaged across the architectures over 10 runs, for $k = [4, 8, 16, 32, 64]$ few-shot training examples.

B Evaluation Metrics

To evaluate our models and have a fair comparison with the introduced benchmarks we use a standard set of metrics for classification tasks such as F1, precision, recall and accuracy.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \qquad (3)$$

$$Prec = \frac{TP}{TP + FP} \qquad (4)$$

$$Recall = \frac{TP}{TP + FN} \qquad (5)$$

$$F1 = \frac{2 * Prec * Recall}{Prec + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \qquad (6)$$

It must be noted that the calculation of the metrics on the expert annotated benchmark is slightly changed as the amount of possible valid annotations can be bigger than 1, decided upon by expert annotators. We use a soft matching approach that allows us calculate the complete set of the evaluation metrics, by counting the annotated example as correct if and only if it matches exactly with at least one $(topic, stance)$ pair in the data-set for the designated sample.

C Transformer variations

The PLMs are taken from the set of *roberta-base, roberta-large, xlm-roberta-base, xlm-roberta-large* with $k = [4, 8, 16, 32, 64]$.

D Exact and Soft Matching in expert annotated data-set

Within our experiments, we use two techniques for validating model performance. The exact matching scheme regards only the first annotation of a $(topic, stance)$ pair as correct, within the final expert benchmark, as throughout the annotation process the first position is reserved only for the most relevant and valid pair. However, due to the similarity in the expressed targets of interest and their intertwined representation within social media sentences, we also employ a soft matching scheme, where a prediction is considered correct if it matches with any $(topic, stance)$ pair present for the designated example within this data-set. Mathematically this can be formalised like the following.

$$match_{exact} = \mathbf{1}(pred_i = \arg\max_{r_i} y_i) \qquad (7)$$

$$match_{soft} = \mathbf{1}(|pred_i \cap y_i| > 0) \qquad (8)$$

Here $\arg\max_{r_i}$ designates the most relevant $(topic, stance)$ pair for the example i, with annotations $y_i \in Y$ and $\max|Y| = 3$ per example.

E Additional Results

	Target: Anthony S. Fauci, M.D.			
	roberta-base TOGA examples = 64	BERT	BERT-NS	BERT-DAN
Accuracy	**0.968**	0.817	0.820	0.830
F1	**0.984**	0.818	0.821	0.832
	Target: Keeping Schools Closed			
	roberta-base TOGA examples = 64	BERT	BERT-NS	BERT-DAN
Accuracy	**0.972**	0.772	0.780	0.758
F1	**0.995**	0.755	0.753	0.717
	Target: Stay At Home Orders			
	roberta-base TOGA examples = 64	BERT	BERT-NS	BERT-DAN
Accuracy	**0.969**	0.843	0.832	0.833
F1	**0.985**	0.800	0.784	0.787
	Target: Wearing a Face Mask			
	roberta-base TOGA examples = 64	BERT	BERT-NS	BERT-DAN
Accuracy	**0.981**	0.810	0.840	0.840
F1	**0.983**	0.803	0.833	0.825

Table 4: Analysis of the best stance model configuration per target topic compared to the proxy benchmark from (Glandt et al., 2021)

		Exact			Soft		
		Match %	Krippendorff's alpha	Cohen's Kappa	Match %	Krippendorff's alpha	Cohen's Kappa
Annotator 1&2	Topics	0.755	0.732	0.722	0.895	0.893	0.880
	Stance	0.707	0.678	0.641	0.675	0.599	0.578
Annotator 3&4	Topics	0.447	0.398	0.387	0.600	0.568	0.552
	Stance	0.973	0.966	0.958	0.843	0.767	0.749

Table 5: Inter Annotator Agreement metrics within each expert annotation group on the expert annotated data-set

			Face Masks			Fauci			School Closures			Stay at Home Orders		
			Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
robert-base	examples = 4	Random	0.661	0.753	0.704	0.854	0.383	0.528	0.532	0.703	0.606	0.642	0.774	0.683
		TOGA	**0.700**	**0.851**	**0.768**	**0.937**	**0.496**	**0.649**	**0.550**	**0.757**	**0.637**	**0.635**	**0.857**	**0.729**
	examples = 8	Random	0.732	0.903	0.809	0.967	0.938	0.952	**0.956**	0.873	0.913	0.849	0.812	0.830
		TOGA	**0.787**	**0.958**	**0.864**	**0.988**	**0.941**	**0.964**	0.948	**0.936**	**0.942**	**0.906**	**0.798**	**0.849**
	examples = 16	Random	0.978	0.922	0.949	**0.952**	0.981	0.966	0.979	0.990	0.984	0.955	0.934	0.944
		TOGA	**0.989**	**0.926**	**0.957**	0.945	**0.995**	**0.969**	**0.988**	**0.991**	**0.990**	**0.950**	**0.969**	**0.960**
	examples = 32	Random	0.939	0.994	0.966	0.965	0.963	0.964	0.981	0.968	0.974	0.942	0.964	0.953
		TOGA	**0.943**	**0.994**	**0.968**	**0.967**	**0.977**	**0.972**	**0.997**	**0.965**	**0.981**	**0.954**	**0.971**	**0.963**
	examples = 64	Random	0.977	0.985	0.981	0.999	0.982	0.990	0.999	0.988	0.993	0.971	0.992	0.978
		TOGA	**0.984**	**0.982**	**0.983**	**0.999**	**0.983**	**0.991**	**0.999**	**0.990**	**0.995**	**0.974**	**0.996**	**0.985**
xlm-robert-base	examples = 4	Random	0.438	0.592	0.503	0.677	0.598	0.635	0.003	1	0.007	0.632	0.224	0.331
		TOGA	**0.598**	**0.604**	**0.601**	**0.694**	**0.624**	**0.781**	**0.003**	**1**	**0.007**	**0.687**	**0.396**	**0.503**
	examples = 8	Random	0.392	0.773	0.520	0.762	**0.642**	0.697	**0.782**	0.691	0.734	0.668	0.497	0.570
		TOGA	**0.443**	**0.798**	**0.570**	**0.794**	0.635	**0.706**	0.761	**0.772**	**0.766**	**0.683**	**0.517**	**0.589**
	examples = 16	Random	0.802	0.912	0.853	0.885	**0.791**	0.835	0.912	0.899	0.905	0.753	0.901	0.820
		TOGA	**0.824**	**0.937**	**0.877**	**0.918**	0.778	**0.843**	**0.933**	**0.914**	**0.924**	**0.747**	**0.920**	**0.825**
	examples = 32	Random	0.974	0.868	0.918	0.905	0.962	0.933	**0.978**	0.959	0.968	0.913	0.985	0.948
		TOGA	**0.991**	**0.885**	**0.935**	**0.911**	**0.986**	**0.947**	0.975	**0.977**	**0.976**	**0.917**	**0.987**	**0.951**
	examples = 64	Random	**0.942**	0.957	0.949	0.967	0.962	0.964	**0.998**	0.912	0.953	0.951	0.943	0.947
		TOGA	0.931	**0.994**	**0.961**	**0.980**	**0.979**	**0.979**	**0.997**	**0.937**	**0.966**	**0.965**	**0.967**	**0.966**

Table 6: Few-shot finetuning experimentation on the proxy data from (Glandt et al., 2021) done with $examples = [4, 64]$ per class with and without the use of ***TOGA*** for generating weakly supervised examples

			Averaged Accuracy	Weight-Averaged F1
roberta-base lin-prob	examples = 4	Random	0.398	0.423
		TOGA	**0.471**	**0.491**
	examples = 8	Random	0.513	0.491
		TOGA	**0.584**	**0.576**
	examples = 16	Random	0.601	0.617
		TOGA	**0.639**	**0.651**
	examples = 32	Random	0.732	0.744
		TOGA	**0.779**	**0.786**
	examples = 64	Random	0.806	0.822
		TOGA	**0.858**	**0.857**
roberta-large lin-prob	examples = 4	Random	0.289	0.358
		TOGA	**0.323**	**0.396**
	examples = 8	Random	0.404	0.458
		TOGA	**0.468**	**0.507**
	examples = 16	Random	0.553	0.512
		TOGA	**0.564**	**0.588**
	examples = 32	Random	0.581	0.574
		TOGA	**0.634**	**0.613**
	examples = 64	Random	0.776	0.801
		TOGA	**0.819**	**0.820**
xlm-roberta-base lin-prob	examples = 4	Random	0.307	0.408
		TOGA	**0.346**	**0.459**
	examples = 8	Random	**0.358**	0.367
		TOGA	0.274	0.372
	examples = 16	Random	0.480	0.524
		TOGA	**0.546**	**0.581**
	examples = 32	Random	0.723	0.718
		TOGA	**0.760**	**0.763**
	examples = 64	Random	0.804	0.832
		TOGA	**0.864**	**0.865**
xlm-roberta-large lin-prob	examples = 4	Random	**0.331**	0.325
		TOGA	0.280	**0.374**
	examples = 8	Random	**0.389**	**0.478**
		TOGA	0.378	0.476
	examples = 16	Random	0.485	0.477
		TOGA	**0.523**	**0.524**
	examples = 32	Random	0.691	0.688
		TOGA	**0.732**	**0.734**
	examples = 64	Random	0.787	0.801
		TOGA	**0.816**	**0.816**

Table 7: Few-shot fine-tuning experimentation with frozen PLM (linear-probing) on the proxy data from (Glandt et al., 2021) done with $examples = [4, 64]$ per class with and without the use of ***TOGA*** for generating weakly supervised examples

Target	to-label	unlabeled
Anthony S. Fauci, M.D.	2,085	2,443
Keeping Schools Closed	1,479	2,703
Stay at Home Orders	1,717	15,488
Wearing a Face Mask	1,921	9,006
All	7,122	29,640

Table 8: Distribution of examples per target topic in the proxy dataset (Glandt et al., 2021)

Topic	Corr	(p-value)
Behaviour change	0.286	(0.49)
Concern, the economy	-0.762	(0.03)
Concern, family	-0.024	(0.96)
Concern, hospitals	-0.167	(0.69)
Concern, society	-0.095	(0.82)
Concern, crime	-0.19	(0.65)
Conspiracy beliefs	0.024	(0.96)
Democratic rights	0.119	(0.78)
Fatigue	-0.619	(0.10)
Knowledge	0.548	(0.16)
Misinformation	0.452	(0.26)
Support of public protests	0.0	(1.0)
Support in restrictions	0.286	(0.49)
Trust in government	0.833	(0.01)
Trust in NHA	-0.143	(0.74)
Trust in scientists	-0.071	(0.87)
Vaccine hesitancy	-0.238	(0.57)

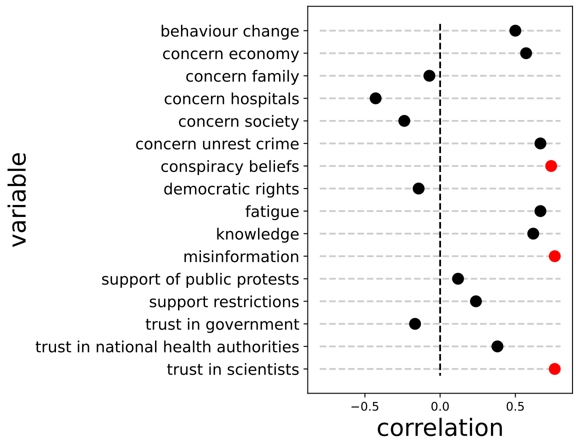
Table 9: Correlations of the Twitter stances with the HOPE survey across all countries.

Topic	Corr	(p-value)
Behaviour change	0.189	(0.07)
Concern, the economy	0.015	(0.88)
Concern, family	-0.073	(0.48)
Concern, hospitals	-0.042	(0.68)
Concern, society	0.013	(0.90)
Concern, crime	-0.071	(0.49)
Conspiracy beliefs	0.080	(0.43)
Democratic rights	0.167	(0.10)
Fatigue	-0.084	(0.42)
Knowledge	-0.036	(0.73)
Misinformation	-0.059	(0.57)
Support of public protests	0.113	(0.28)
Support in restrictions	-0.116	(0.26)
Trust in government	0.162	(0.12)
Trust in NHA	-0.170	(0.10)
Trust in scientists	-0.022	(0.83)
Vaccine hesitancy	-0.177	(0.09)

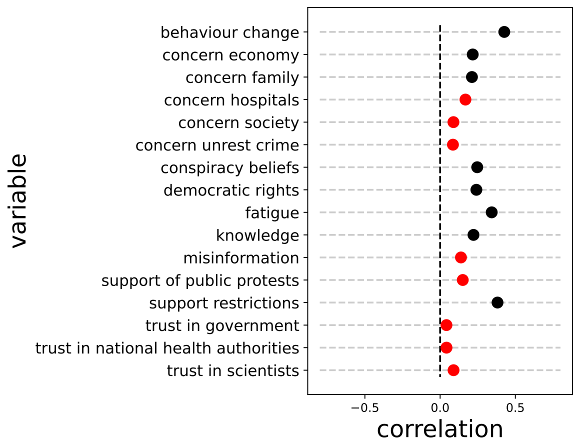
Table 10: Correlations of the Twitter stances with the HOPE survey, breaking into states and counties.

Topic	Correlation	
	Country	State
Concern, the economy	**-0.762**	0.015
Concern, hospitals	-0.166	-0.042
Conspiracy beliefs	0.024	0.080
Misinformation	0.452	-0.059
Support in restrictions	0.286	-0.116
Trust in the government	**0.833**	0.162
Trust in scientists	0.071	-0.022
Vaccine hesitancy	0.238	-0.177

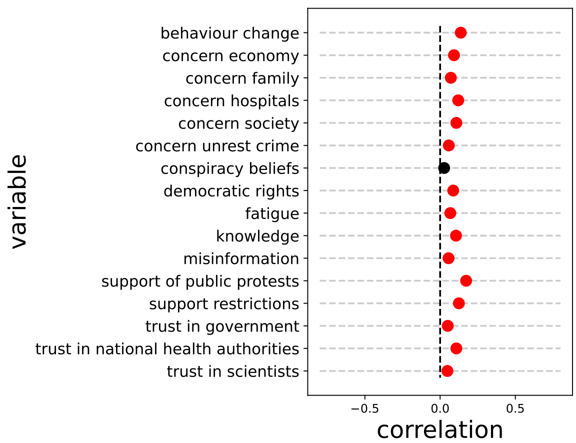
Table 11: Correlations of the Twitter stances with the survey, across countries and states. Items in bold are statistically significant (p-value < 0.05).



(a) Country level



(b) State level



(c) User level

Figure 4: Predictors of vaccine acceptance. Red markers indicate p-value < 0.5.

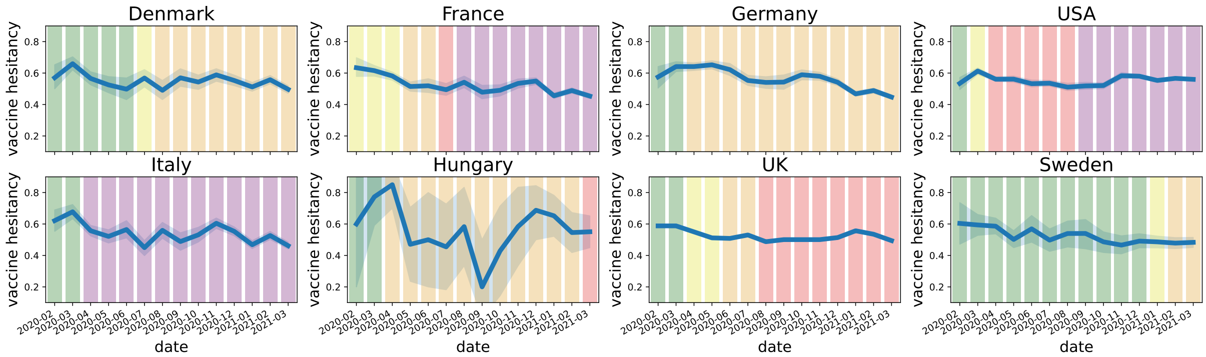


Figure 5: Development in vaccine hesitancy over time across countries. The background colour corresponds to the severity of Covid restrictions related to face masks. Green = no restrictions. Yellow = recommended. Orange = required in some specified shared/public spaces outside the home with other people present, or some situations when social distancing not possible. Red = required in all shared/public spaces outside the home with other people present or all situations when social distancing not possible. Purple = required outside the home at all times regardless of location or presence of other people.

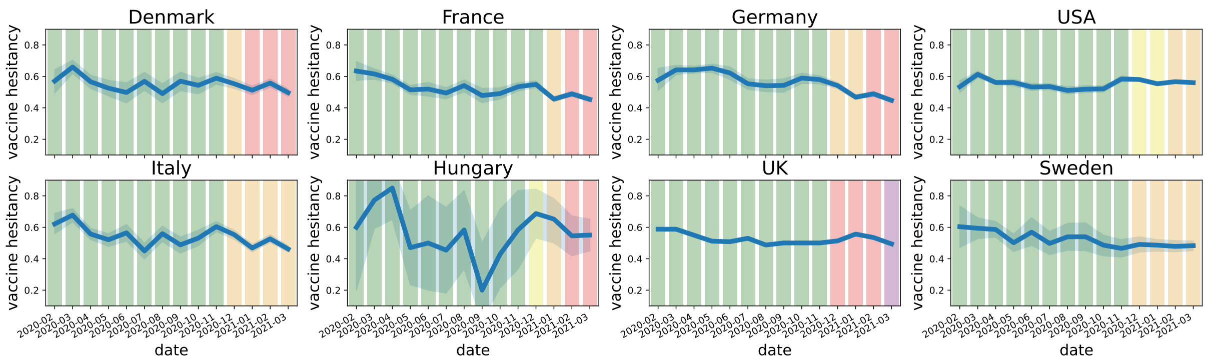


Figure 6: Development in vaccine hesitancy over time across countries. The background colour corresponds to the vaccination policy. Green = no vaccine available. Yellow = availability for ONE of following: key workers/ clinically vulnerable groups / elderly groups. Orange = availability for TWO of following: key workers/ clinically vulnerable groups / elderly groups. Red = availability for ALL of following: key workers/ clinically vulnerable groups / elderly groups. Purple = availability for all three plus partial additional availability (select broad groups/ages).