Mining the Cause of Political Decision-Making from Social Media: A Case Study of COVID-19 Policies across the US States

Zhijing Jin
MPI & ETH Zürich
zjin@tue.mpg.de

Zeyu Peng
MIT Political Science
pzy0337@mit.edu

Tejas Vaidhya
IIT Kharagpur
iamtejasvaidhya@iitkgp.ac.in

Bernhard Schoelkopf
MPI & ETH Zürich
bs@tue.mpg.de

Rada Mihalcea
University of Michigan
mihalcea@umich.edu

Abstract

Mining the causes of political decision-making is an active research area in the field of political science. In the past, most studies have focused on long-term policies that are collected over several decades of time, and have primarily relied on surveys as the main source of predictors. However, the recent COVID-19 pandemic has given rise to a new political phenomenon, where political decision-making consists of frequent short-term decisions, all on the same controlled topic—the pandemic. In this paper, we focus on the question of how public opinion influences policy decisions, while controlling for confounders such as COVID-19 case increases or unemployment rates. Using a dataset consisting of Twitter data from the 50 US states, we classify the sentiments toward governors of each state, and conduct controlled studies and comparisons. Based on the compiled samples of sentiments, policies, and confounders, we conduct causal inference to discover trends in political decision-making across different states.

1 Introduction

Policy responsiveness is the study of the factors that policies respond to (Stimson et al., 1995). One major direction is that politicians tend to make policies that align with the expectations of their constituents, in order to run successful re-election in the next term (Canes-Wrone et al., 2002).

An overview of existing studies on policy responsiveness reveals several patterns, summarized in Table 1. First, most work focuses on the *long-term* setting, where the policies are collected over a span of several decades, e.g., Caughey and Warshaw (2018)'s collection of public opinion surveys and state policymaking data over 1936-2014, and Lax and Phillips (2009)'s collection of public opinion polls and gradual policy changes over 1999-2008. Second, the data sources of existing studies are mostly surveys and polls, which can be time-consuming and expensive to collect (Lax and Phillips, 2012). Third, the resulting data are often of relatively small sizes, for both the number of policies and the number of public opinion.

	Previous Work	This Work
Policy Type	Long-term, gradual (over decades)	Short-term (weekly/monthly)
Policy Sparsity	Less policies on the same topic	Many policies on the same topic across states
Data Source	Surveys	Trillions of tweets
Data Collection	–	NLP & Causality

Table 1: Comparison of the characteristics and paradigms of existing work versus our work.

Different from previous work on long-term policies, our work focuses on the special case of COVID pandemic, during which political leaders make a number of frequent, short-term policies on the same topic: social distancing. Moreover, instead of collecting surveys, we use Twitter to collect public opinion, which is instant, costless, and massive, e.g., trillions of data points. We limit our scope to US policies because the 50 states provide abundant policy data, and a good background for both controlled groups and comparative studies.

We present one of the first efforts to address policy responsiveness for short-term policies, namely the causal impact of public Twitter sentiments on political decision-making. This is distinct from existing studies on COVID policies that mostly explore the impact of policies, such as predicting public compliance (Grossman et al., 2020; Allcott et al., 2020; Barrios and Hochberg, 2020; Gadarian et al., 2021; DeFranza et al., 2020). Specifically, since governors have legislative powers through executive orders, we focus our study on each state governor's decisions and how public opinion towards the governor impacts their decisions. For example, governors that optimize short-term public opinion are more likely to re-open the state even when case numbers are still high.

Our workflow is illustrated in Figure 1. We start by collecting 10.4M governor-targeted COVID

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 288–301
November 7–11, 2021.

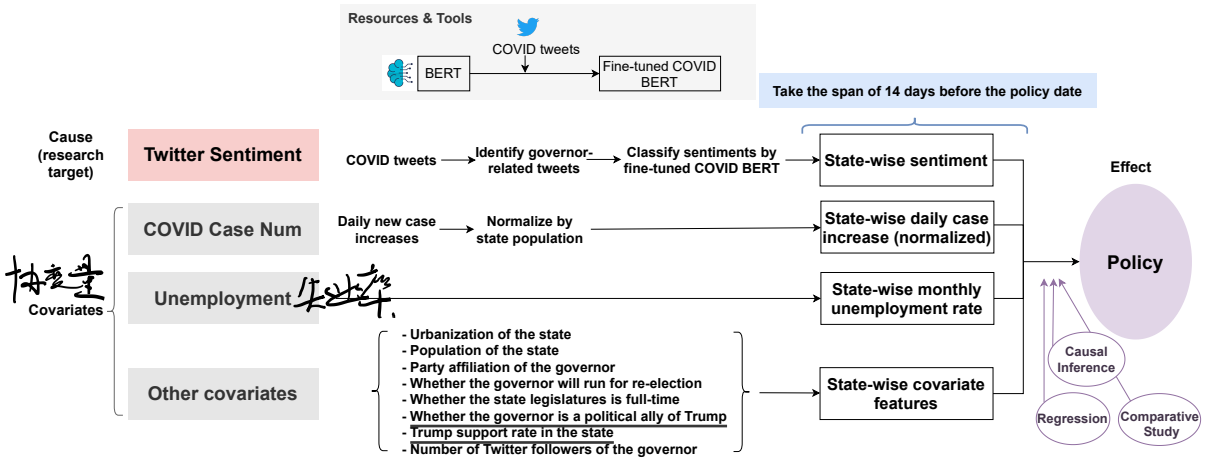


Figure 1: The data collection pipeline and architecture of our system to predict the state-wise COVID policies.

tweets, which we annotate for sentiment with a BERT-based classifier. Next, we annotate 838 social distancing policies and collect data on ten potential confounders such as average daily case increases or unemployment rates. Finally, we conduct multiple analyses on the causal effect of Twitter sentiment on COVID policies. For interpretability, we first use a multivariate linear regression to identify correlations of sentiments and policies, in addition to considering all the confounders. We also use do-calculus (Pearl, 1995) to quantify the causal impact of Twitter sentiment on policies. We also conduct cross-state comparisons, cross-time period analysis, and multiple other analyses.

The main contributions of our work are as follows. First, we compile a dataset of public opinion targeted at governors of the 50 US states with 10.4M tweets. Second, we annotate a dataset of 838 COVID policy changes of all 50 states, along with data of ten confounders of each state. Third, we conduct regression analyses and causal analyses on the effect of Twitter sentiment on policies. Finally, we implement additional fine-grained analyses such as cross-state comparisons, cross-time period analysis, and multiple other analyses.

2 Related Work

Policy Responsiveness. Policy responsiveness (i.e., public opinion $\xrightarrow{\text{causes}}$ policies) is an active research field in political science, where people study how policies respond to different factors (Stimson et al., 1995). Studies show that policy preferences of the state public can be a predictor of future state policies (Caughey and Warshaw, 2018). For example, Lax and Phillips (2009) show that more LGBT tolerance leads to more pro-gay legislation in response. Most policies and public opinion studied in existing literature are often long-term and gradual, taking several decades to observe (Lax and Phillips, 2009, 2012; Caughey and Warshaw, 2018).

Crisis Management Policies. Another related topic is crisis management policies, where most studies focus on the reverse causal problem of our study – how crisis management policies impact public opinion (i.e., policies $\xrightarrow{\text{causes}}$ public opinion). A well-known phenomenon is the rally "round the flag" effect, which shows that during a crisis, there will be an increased short-run public support for the political leader (Mueller, 1970, 1973; Baum, 2002), due to patriotism (Mueller, 1970; Parker, 1995), lack of opposing views or criticism (Brody and Shapiro, 1989), and traditional media coverage (Brody, 1991).

To the best of our knowledge, there is not much research on how public opinion influence policies (i.e., public opinion $\xrightarrow{\text{causes}}$ policies) during a crisis. Our work is one of the few to address this direction of causality.

COVID-19 Policies. There are several different causal analyses related to COVID-19 policies, although different from our research theme. Existing studies focus on how social distancing policies mitigate COVID spread (i.e., policies $\xrightarrow{\text{causes}}$ pandemic spread) (Kraemer et al., 2020), what features in public attitudes impact the compliance to COVID policies (i.e., public attitudes/ideology $\xrightarrow{\text{causes}}$ policy compliance) (Grossman et al., 2020; Allcott et al., 2020; Barrios and Hochberg, 2020; Gadarian et al., 2021), how polices change the public support of leaders (i.e., policy $\xrightarrow{\text{causes}}$ public support). Bol et al. (2021); Ajzenman et al. (2020), how pandemic

characteristics affect Twitter sentiment (Gencoglu and Gruber, 2020), and how political partisanship impacts policies (i.e., partisanship $\xrightarrow{\text{causes}}$ policy designs) (Adolph et al., 2021). However, there is no existing work using public sentiments (e.g., from social media) to model COVID policies.

Opinion Mining from Social Media. Social media, such as Twitter, is a popular source to collect public opinions (Thelwall et al., 2011; Paltoglou and Thelwall, 2012; Pak and Paroubek, 2010; Rosenthal et al., 2015). Arunachalam and Sarkar (2013) suggest that Twitter can be a useful resource for governments to collect public opinion. Existing usage of Twitter for political analyses mostly targets at election result prediction (Beverungen and Kalita, 2011; Mohammad et al., 2015; Tjong Kim Sang and Bos, 2012), and opinion towards political parties (Pla and Hurtado, 2014) and presidents (Marchetti-Bowick and Chambers, 2012). To the best of our knowledge, this work is one of the first to use Twitter sentiment for causal analysis of policies.

3 Governor-Targeted Public Opinion

To investigate the causality between public opinion and each state governor's policy decisions, we first describe how we mine public opinion in this Section; we then describe the process we use to collect policies and other confounders in Section 4.

We collect governor-targeted public opinion in two steps: (1) retrieve governor-related COVID tweets (Section 3.1), and (2) train a sentiment classification model for the COVID tweets and compile sentiments towards governors (Section 3.2).

3.1 Retrieve Governor-Related COVID Tweets

We use the COVID-related tweet IDs curated by Chen et al. (2020).[1] Chen et al. (2020) identified these tweets by tracking COVID-related keywords and accounts. We provide the list of keywords and accounts they used in Appendix A.1. We hydrate the tweet IDs to obtain raw tweets using an academic Twitter Developer account. This process took several months to complete, and resulted in a dataset of 1.01TB. The retrieved 1,443,871,617 Tweets span from January 2020 to April 2021.

Since this study focuses on governor's policy decision-making process, we focus on the public opinion that are more directly related to the governors. Specifically, we focus on tweets that tagged, replied to, or retweeted state governors. We obtain 10,484,084 tweets by this filter. On average, each of the 50 states has about 209K tweets that address the state governor. The rationale of this filter is that the governors and their teams are likely to have directly seen (a portion of) these tweets, since they showed up in governor's Twitter account.

3.2 Classify Sentiments towards Governors

Existing studies on COVID Twitter sentiment analysis (Manguri et al., 2020; Kaur and Sharma, 2020; Vijay et al., 2020; Chakraborty et al., 2020; Singh et al., 2021) mostly use TextBlob (Loria, 2018), or some simple supervised models (Machuca et al., 2021; Kaur et al., 2021; Mansoor et al., 2020).

For our study, we use the state-of-the-art BERT model pretrained on COVID tweets by Müller et al. (2020).[2] We finetune this pretrained COVID BERT on the Twitter sentiment analysis data from SemEval 2017 Task 4 Subtask A (Rosenthal et al., 2017). Given tweets collected from a diverse range of topics on Twitter, the model learns a three-way classification (positive, negative, neutral). In the training set, there are 19,902 samples with positive sentiments, 22,591 samples with neutral sentiments, and 7,840 samples with negative sentiments.

We tokenize the input using the BERT tokenizer provided by the Transformers Python package (Wolf et al., 2020). We add [CLS] and [SEP] tokens at start and end of the input, respectively. The input is first encoded by the pretrained COVID BERT. Then, we use the contextualized vector C of the [CLS] token as the aggregate sentence representation. The model is finetuned on the classification task by training an additional feed-forward layer $\log(\text{softmax}(CW))$ that assigns the softmax probability distribution to each sentiment class.

Prior to training, we preprocess the tweets by deleting the retweet tags, and pseudonymising each tweet by replacing all URLs with a common text token. We also replace all unicode emoticons with textual ASCII representations. During training, we use a batch size of 32 and fine-tune for 5 epochs. We use a dropout of 0.1 for all layers, and the Adam optimizer (Kingma and Ba, 2017) with a learning rate of 1e-5. Additionally, due to the specific nature of our classification task (i.e., mining opinion towards the governor), we add a post-processing step to classify a tweet as supportive of a governor

[1] COVID-related Tweet IDs: https://github.com/echen102/COVID-19-TweetIDs

[2] https://huggingface.co/digitalepidemiologylab/covid-twitter-bert-v2

	Positive	Neutral	Negative
Percentage	15.8%	36.5%	47.7%
Length	15.51	12.21	16.39
Topics	we, support, thank, great, governors, covid, action	people, masks, covid, cases, state, today, total	cases, state, covid, close, deaths, people, trump
4-Grams	- great governors responded executive - responded executive action promptly - quickly , support americans	- positive patients nursing homes - governors ordered covid positive - today 's update numbers	- covid patients nursing homes - america 's governors forced - covid patients nursing homes
Example	"I am a small business owner, we kept health insurance for the furloughed staff of my two restaurants, month after month, even while one restaurant was closed and the other only has limited service. Why? Because I have a conscience. We are in a pandemic."	"Today: @GovInslee 3 pm news conference on WA's coronavirus response. Inslee to be joined by state schools chief. Your daily #covid19 updates via @seattletimes"	"And the politicians that are doing the conditioning are out, maskless, celebrating with their family and friends... @GavinNewsom Glad I never once fell for it. Covid-19 was always just a power-grab for politicians"

Table 2: Label distribution (Percentage), average number of words per tweet (Length), topics extracted by LDA topic modeling (Blei et al., 2003), top 4-grams, and examples of positive, neutral, and negative tweets.

(i.e., positive) if the tweet retweets a tweet from the governor's official account.

Model Performance. We evaluate our model accuracy on two test sets. First, on the test set of SemEval 2017, our finetuned model achieves 79.22% accuracy and 79.29% F1. Second, we also evaluate our model performance on our own test set. Since the features of general tweets provided in SemEval 2017 might differ from COVID-specific tweets, we extracted 500 random tweets from the Twitter data we collected in Section 3.1. We asked a native English speaker in the US to annotate the Twitter sentiment with regard to the state governor that the tweet addresses. The annotator has passed a small test batch before annotating the entire test set.

We use the TextBlob classifier as our baseline, since it is the most commonly used in existing COVID Twitter sentiment analysis literature. On our test set's three-way classification, the TextBlob baseline has 23.35% accuracy and 16.67% weighted F1. Our finetuned BERT classifier has 60.23% accuracy and 62.31% weighted F1. Detailed scores per class is in Appendix A.3. When applying the sentiment classifier, we care more about whether the average sentiment over a time period is accurate, so we also turn the test set into groups of tweets each containing 20 random samples. The average mean squared error (MSE) for the average sentiment of each group is 0.03889 for the BERT model, and 0.22749 for the TextBlob model. We apply the finetuned COVID BERT classifier on the governor-related tweets we extracted previously. As listed in Table 2, among 10.4M tweets, 15.8% are positive, 36.5% neutral, and 47.7% negative.[3]

[3]Note that label imbalance is commonly observed on Twitter data (Guerra et al., 2014).

We use Latent Dirichlet Allocation (LDA) topic modeling (Blei et al., 2003) to extract key topics of each category. Typical topic words in positive tweets include "we," "support," "thank," "great," and "governors," while negative tweets tend to mention more about "america's governors forced ..." and support Trump, perhaps Trump's tweets on "liberation."

4 Collection of Policies and Confounders

We focus on state-wide social distancing policies, and collect 838 social distancing policies from 50 states over the period January 2020 – April 2021 (described in Section 4.1).

Since we want to focus on the causal effect of public sentiment on policy, we must control for possible confounding factors. In particular, case numbers and unemployment rates are potentially the most important confounders, the collection of which is introduced in Section 4.2. In addition, we also collect eight other potential confounders suggested by political science experts (described in Section 4.3). The collection process is illustrated in Figure 1.

4.1 Social Distancing Policy Annotation

We annotate the social distancing policies related to COVID for each of the 50 states in the US. For each state, the annotators are asked to go through the entire list of COVID-related executive orders from January 2020 to April 2021. In cases where the states do not use executive orders for COVID regulations, we also consider proclamations and state guidance on social distancing.

The policies are rated on a scale of 0 (loosest) - 5 (strictest). We provide guidance as to the level of strictness that each number indicates, as detailed

in Appendix A.2. Four annotators are asked to conduct the ratings. Since the annotation is very tedious, taking up to 3 hours per state, we do not conduct double annotations. Instead, given our original annotations (for which we score each policy based on its official legal document in PDF), we did a quick second pass by confirming that our scores roughly match the succinct 1∼2-sentence textual summary of each policy provided by the Johns Hopkins Coronavirus Resource Center.[4]

4.2 Key Confounders: State-Level Case Numbers and Unemployment Rates

We collect COVID daily new confirmed case numbers from the open-source COVID database[5] curated by the Kaiser Family Foundation. For a fair comparison across states, we normalize the case numbers by the population of the state. We retrieve the seasonly adjusted data of monthly unemployment rates for each state from the U.S. Bureau of Labor Statistics.[6]

4.3 Additional Confounders

For additional confounders, we collect both state data as well as governor features.

State Features. For state features, we collect the population[7] and urbanization rate from US 2010 Census (Census Bureau, 2012).[8] In addition, we also collect the last US presidential election returns of each state.[9] Note that it is necessary to use pre-policy data, so we collect the presidential election returns from 2016 but not from 2020. For the presidential election returns, we obtain the percentage of votes for Donald Trump to indicate Trump's support rate.

Governor Features. For each governor, we collect their party affiliation, whether the governor will run for the next gubernatorial election,[10] and whether the state legislatures are full-time or not, collected from National Conference of State Legislatures.[11] In addition, we also annotate whether the governor is a political ally of Trump or not. We conduct the annotation based on the background and past news reports of each governor. For corner cases, we quote additional evidence in our annotation, e.g., for republican governors who do not support Trump, and democratic governors who support Trump. We also collect the number of Twitter followers for each governor, since it might be correlated with how much attention the governor pays to the twitter reactions.

Table 3 lists the statistics of the confounder data we collected.

Numerical Features			
	Mean (±std)	Min	Max
Daily case increase (%)	0.02 (±0.02)	0.0	0.45
Unemployment rate (%)	5.51 (±3.25)	2.0	29.5
Urbanization (%)	73.58 (±14.56)	38.7	95
Population (M)	12.94 (±45.68)	0.57	325.38
Trump's support rate (%)	48.29 (±11.93)	4	68
# Twitter followers (K)	237 (±458)	7	2596
Binary Features			
	Yes	No	
Gov is republican	26	24	
Will run for re-election	39	11	
Full-time legislatures	10	40	
Trump's political ally	22	28	

Table 3: Statistics of the ten confounders collected for policy prediction task.

5 Mining Decisive Factors of COVID Policies

Since we are interested in discovering the key factors that changes the decisions of policy-makers, we focus on the change of policies (e.g., changing from complete close down to reopening K-12 schools) rather than absolute values of the policy strictness. For each policy in state s on date t, we calculate the change Δpolicy as the difference of this policy from the previous policy that was issued.

Since sentiment may change rapidly and many policies are updated frequently during COVID, for each policy change Δpolicy, we focus on the average sentiment over the time span $(t - \Delta t, t)$ from Δt days prior to the policy date t. Here, we set $\Delta t = 14$ since many epidemiology reports are based on 14-day statistics, e.g., the 14-day notification rate.

When building the policy prediction model, we also need to account for confounders. For the confounders, most are static over time for a given state,

[4] Social distancing policy summaries: https://coronavirus.jhu.edu/data/state-timeline

[5] COVID case number data: https://github.com/KFFData/COVID-19-Data

[6] Monthly unemployment data: https://www.bls.gov/web/laus/ststdsadata.zip

[7] Population data: https://www.census.gov/programs-surveys/decennial-census/data/tables.2010.html

[8] Urbanization data: https://www.icip.iastate.edu/tables/population/urban-pct-states.

[9] Presidential election return data: https://www.nytimes.com/elections/2016/results/president

[10] For simplicity, we collect the pre-COVID data at the time point of January 2020, and do not consider the change of governorships in two states in early 2021.

[11] https://www.ncsl.org/

except for the daily case increases and the unemployment rates that change over time, for which we take the average over the 14-day time span.

Based on the data above, we seek to answer the following questions: (Q1) What variables are indicative of policy changes?, and (Q2) What causal impact does sentiment have on the policies?

5.1 Q1: What Variables Are Indicative of Policy Changes?

To aim for interpretability, we choose a multivariate linear regression as our model, which is commonly used in political science literature on COVID policies (Grossman et al., 2020; Allcott et al., 2020; Barrios and Hochberg, 2020; Gadarian et al., 2021). Specifically, we model the policy change Δpolicy as a function of all variables, including our main focus – Twitter sentiments – and all the confounders, which form in total 11 variables.[12]

Sentiment, Case Numbers, Unemployment Are Important. The first experiment is to compare how well different combinations of input variables fit the policy change. We use mean squared error (MSE) as the measure of model capability.

When taking into consideration all variables, the model has an MSE score of 0.368. As a further step, we test whether a smaller number of inputs can achieve similar results. We find that when only taking three variables as inputs, the MSE is 0.369, which is 0.001 from the model taking in all variables. Among all combinations of three variables, the proposed three key variables, sentiment, case numbers, and unemployment rates, achieve the best performance of 0.369.

Note that it is reasonable that with rational decision-making, politicians consider the case numbers and unemployment rates when making COVID policies. The focus of this study is to show the *additional* effect of sentiment, the role of which is not explicitly pointed out in previous COVID policy research.

The Role of Non-Sentiment Variables. First, given the presence of the sentiment variable in the model, we test the additional effect of non-sentiment variables. As shown in Table 4, case numbers and unemployment rate both lead to non-trivial improvement of the models, and unemployment is more important.

The Role of Sentiment. Second, we look into the role of sentiment. We take the optimal 11-variable, 3-variable, and 2-variable models, and conduct ablation studies to inspect how much does sentiment contribute exclusively in Table 5.

[12]For each input variable, we first normalize by adjusting mean to zero and standard deviation to 1.

Additional Non-Sentiment Variables	MSE (↓)
Sentiment-only	0.618
+ Case	0.532
+ Unemp	0.407
+ Case, Unemp	0.369
+ Case, Unemp, Others	0.368

Table 4: The MSE of models taking as input the additional non-sentiment variables, such as case increases (Case), unemployment (Unemp), and other confounders (Others).

We show that for each model, sentiment has a crucial impact of more than 0.032 on the model performance. Note that in linear regression, we do not need to explicitly disentangle the correlations within sentiments and other confounders – in Table 5, the effect of sentiment is demonstrated in addition to fitting all other variables that may contain correlations.

Model	MSE (↓)
11-Variable model	0.368
−Senti	Deterioration of 0.032
3-Variable model	0.369
−Senti	Deterioration of 0.032
2-Variable model	0.407
−Senti	Deterioration of 0.034

Table 5: Ablation study of sentiment for the optimal 11-, 3-, 2-variable models. Note that the 11-variable model is the full model taking in all variables.

5.2 Q2: What Causal Impact Does Sentiment Have on the Policies?

In the previous section, we investigated the most indicative variables of policies. The experiments indicate how important each variable is to the regression target, i.e., how well they serve as a predictor, although such *correlation* does not necessarily capture *causation*. In this section, we are interested in the causal impact of sentiment on policies, and we use causal inference methods to quantify the impact.

Formulation by Do-Calculus. Formally, we are interested in the effect of a cause X (i.e., Twitter sentiment) on the outcome Y (i.e., policy change) in the presence of the confounder Z (i.e., case numbers, unemployment, etc.), as shown in Figure 2.

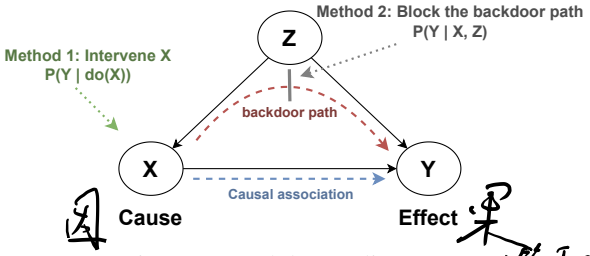


Figure 2: Backdoor Adjustment.

To formulate the causal impact, Pearl (1995) defines a language for causality called do-calculus, by which the causal impact of X on Y is formulated as the interventional distribution:

$$P(Y|\text{do}(X)) \,, \tag{1}$$

where $\text{do}(X)$ refers to an intervention on the cause X.

Note that the interventional distribution $P(Y|\text{do}(X))$ may be different from the observational distribution $P(Y|X)$ in the presence of the confounder Z. Specifically, in the above Figure 2, there are two ways how X correlates with Y. The first is the causal path $X \rightarrow Y$, and the second is the backdoor path $X \leftarrow Z \rightarrow Y$.

There are two ways to account for the backdoor path: Method 1 needs to intervene on X, e.g., create a counterfactual situation where all confounders are the same but the Twitter sentiment can be set to negative vs. positive. In our study of Twitter opinion on COVID policies, this is not a feasible experiment to conduct, due to the fundamental problem of causal inference (Rubin, 1974; Holland, 1986) (namely, for each sample i, we are usually only able to observe one value of X but not both). The other method, backdoor adjustment, circumvents the problem, which will be introduced in the following.

Backdoor Adjustment. The key challenge in the above causal inference is that we need to account for the confounder Z. Backdoor adjustment (Pearl, 1995) presents an approach to estimate the causal impact of X on Y by using only *observational* data. Basically, we need to block all backdoor paths by conditioning on nodes that can break the unwanted connections between X and Y. Moreover, these nodes should not contain any descendants of X. In our case, we condition on the confounder Z, and turn the interventional distribution into the observational distribution:

$$P(Y|\text{do}(X)) = \sum_Z P(Y|X,Z)P(Z) \,. \tag{2}$$

The causal impact of X (i.e., positive or negative sentiment) on Y (i.e., policy change) becomes

$$\begin{aligned}\beta &= \mathbb{E}[Y|\text{do}(X=1)] - \mathbb{E}[Y|\text{do}(X=-1)] \\ &= \sum_Z (\mathbb{E}[Y|X=1,Z] - \mathbb{E}[Y|X=-1,Z])P(Z) \\ &= \mathbb{E}_Z[\mathbb{E}[Y|X=1,Z] - \mathbb{E}[Y|X=-1,Z]] \,.\end{aligned} \tag{3}$$

Results. We apply Eq. (3) to all states using a 10-dim vector Z that encodes all confounders.[13] Then we rank the states by β values, which represents the causal impact of sentiment on the state policies.

Top 5 States with Large β		**Top 5 States with Small $\lvert\beta\rvert$**	
State	β Value	State	β Value
Colorado	4.292	Arizona	0.053
Massachusetts	1.157	West Virginia	0.030
Florida	1.124	Pennsylvania	0.023
Texas	1.095	Nebraska	-0.001
South Dakota	1.057	Alabama	-0.065

Table 6: Top five states with the largest β values, and the β values that are closest to zero.

In Table 6, we show the top five states with highest β values, and five states with β values that are the closest to zero. The higher the β value, there exists more alignment between people's sentiment and the state policy strictness in the state.

There are some associations between our results and real-world patterns. For instance, among the top five states in Table 6, Colorado's high β value reflects its Democratic governor's large net favorable rating compared to the Republican politicians.[14] Massachusetts also has a high governor approval rate, and most people support the COVID policies. The three Republican states, South Dakota, Texas, and Florida, also have high β, but they are in a different scenario. The loose policies in all these states are in line with general sentiment across the states to refuse restrictions.

6 Fine-Grained Analyses

6.1 Early-Stage vs. Late-Stage Decisions

Since the COVID pandemic is an unprecendented situation, it is likely that in early stages of the pandemic, politicians tend to rely on their prejudgements, and as time goes on, they form a better understanding of the situation and adjust their reaction towards the public opinion. We compare

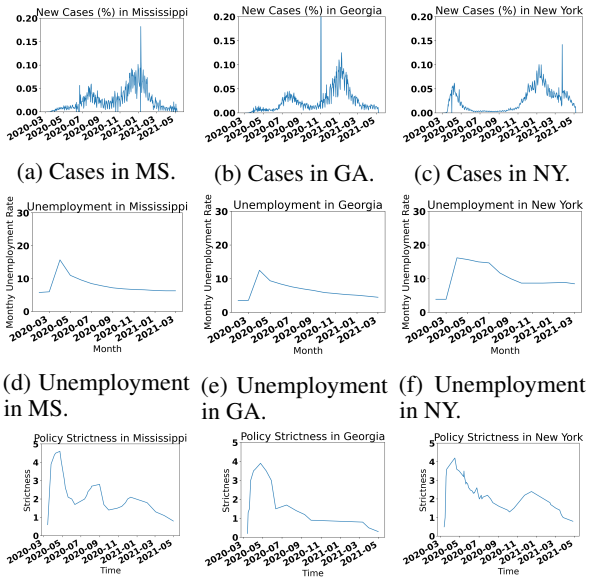
[13] Due to length restrictions, please refer to the arXiv version of our paper for additional implementation details of the backdoor adjustment.

[14] For example, see this poll result by Colorado Poll reported by Denver Post.

the causal impact of sentiment on policies in the first three months of the outbreak (i.e., from March to June 1, 2020) and afterwards (i.e., from June 1, 2020 to now). Table 7 shows that the states with the most changes in β are Montana, Washington, Georgia, Tennessee, and Indiana.

State	Change in β before and after June 1
Montana	+9.39
Washington	+4.03
Georgia	+3.15
Tennessee	+2.94
Indiana	+2.53

Table 7: Top 5 states with the most change in the causal impact of sentiment on policies from March to June 1, 2020, versus from June 1, 2020 to April, 2021.



(a) Cases in MS. (b) Cases in GA. (c) Cases in NY.

(d) Unemployment in MS. (e) Unemployment in GA. (f) Unemployment in NY.

(g) Policy of MS. (h) Policy of GA. (i) Policy of NY.

Figure 3: Comparative study of states. MS and GA is a pair of states with the most similar confounders, and NY is an irrelevant state to contrast how different MS and GA are from other states. Note that unemployment data is only available until March 2021.

6.2 Cross-State Comparison

For cross-state comparison, we identify states that are similar in terms of the confounders, and then compare how different policies are a result of different public sentiments. For simplicity, we consider the two most important confounders, case numbers and unemployment rates. We evaluate the similarity matching on the two time series across different states by the dynamic time warping algorithm (Berndt and Clifford, 1994), and extract state pairs that are the most similar in terms of the confounders.

In Figure 3, we show an example pair of states, Mississippi (MS) and Georgia (GA), which have highly similar case numbers and unemployment rates at most time steps. Note that we use the New York (NY) state to show in contrast how the above pair is different from another unrelated state.

In the comparative study of MS and GA, they can be considered as counterfactuals for each other. In their policy curves, the policy strictness in MS responds to the COVID case numbers (e.g., the policies are stricter on the rising slope of case numbers), but the policies in GA remain loose even during the rising trends in July – August 2020, and November 2020 – January 2021. We look into the sentiment differences across the two states: For example, during November 2020 – January 2021, GA experienced a very low average sentiment of -0.58 in the [-1, 1] scale, whereas MS experienced a milder sentiment of -0.04. By the controled comparison, the more negative sentiment is the potential cause for looser policies in GA.

7 Additional Discussions

Fine-Grained Opinions behind the Sentiments. To further interpret why positive tweets usually lead to stricter social distancing policies (and negative tweets lead to looser policies), we look into the correlation of Twitter sentiment and the user's opinion towards social distancing policies. Note that usually it is not easy to directly get an unsupervised intent classifier on COVID specific tweets. Hence, we ask the annotators to classify the opinion on social distancing for the 500 tweets in our test set as supportive, against, and not related to social distancing. Among the tweets about social distancing with positive sentiment, 95.13% support social distancing. Among the tweets about social distancing with negative sentiment, 69.38% are against social distancing and ask for the reopening of the state.

Additional Analyses. We put our additional analyses in Appendix B, including correlation across all variables, and alternative causal analysis models such as difference-in-differences (Abadie, 2005), and continuous-valued propensity score matching (Hirano and Imbens, 2004; Bia and Mattei, 2008).

Limitations. There are several limitations of this study. For example, a common limitation of many

causal inference settings is the uncertainty of hidden confounders. In our study, we list all the variables that we believe should be considered, but future studies can investigate the effect of other confounders.

Another limitation is the accuracy of the Twitter sentiment classifier. Since the Twitter sentiment during COVID is very task-specific, modeling the sentiments can be very challenging. For example, our model often misclassifies "increased positive cases" as a positive sentiment. Another challenge is that some tweets refer to a url. These cases are difficult to deal with, and might be worth more detailed analyses in future studies.

In the data setting, one limitation is that for causal inference, modeling the whole time series is extremely challenging, so we empirically take the 14-day time span, which is a commonly used time span for many other COVID measures.

Future Work. This work is the first work to use NLP and causal inference to address policy responsiveness, and we explicitly measure the alignment of government policies and people's voice. This signal can be very important for the government and decision-makers.

In future work, a similar approach can be used together with other variables (e.g., economic growth, participation in health/vaccination campaigns, well-being) to determine to which extent such people-government alignment relates to societal outcomes.

8 Conclusion

In this paper, we conducted multi-faceted analyses on the causal impact of Twitter sentiment on COVID policies in the 50 US states. To enable our study, we compile a large dataset of over 10 million governor-targeted COVID tweets, we annotate 838 state-level policies, and we collect data ten potential confounders such as daily COVID cases and unemployment rates. We use a multivariate linear regression and do-calculus to quantify both the correlation of Twitter sentiment as well as its causal impact on policies, in the presence of other confounders. To our knowledge, this is one of the first studies to utilize massive social media data on crisis policy responsiveness, and lays the foundation for future work at the intersection of NLP and policy analyses.

Our code and data are publicly available at https://github.com/zhijing-jin/covid-twitter-and-policy.

Acknowledgements

We thank Kevin Jin for insightful opinions that motivated this work. We thank Jingwei Ni, Yiwen Ding, and Lea Künstler for annotating the state policies. We thank Yiwen Ding for annotating the Twitter test set, and performing the experiment of continuous-valued propensity score matching shown in the Appendix. We thank Di Jin for helping with computational resources. We thank the labmates in the LIT Lab at University of Michigan, especially Ian Stewart, MeiXing Dong, and Laura Biester for constructive suggestions and writing advice.

This material is based in part upon worked supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B; by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645; by the Precision Health Initiative at the University of Michigan; and by the John Templeton Foundation (grant #61156).

Ethical Considerations

Use of Data. For the data used in this study, the COVID-related tweets are a subset of the existing dataset provided by Chen et al. (2020). Following the data regulations of Twitter, we will not publicize the raw tweet text. If necessary, we can provide the list of tweet IDs to future researchers. For the policy strictness we annotated, we will open-source it since it is public information that can benefit societies affected by the pandemic, and has no privacy or ethical issues. For other confounding variables, the data are also public information.

Potential Stakeholders. This research can be used for policy-makers or political science researchers. The research on causality between public opinion and political decision-making helps make policies more interpretable. One potential caveat is that there might be parties that maliciously manipulate sentiments on Twitter to affect politicians. A mitigation method is to control the flow of misinformation, terrorism and violent extremism on social media. The ideal use of the study is to reflect the process how a democracy system surveys the opinion from people, and makes policies that best balances people's long-term and short-term interests.

References

Alberto Abadie. 2005. Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19.

Christopher Adolph, Kenya Amano, Bree Bang-Jensen, Nancy Fullman, and John Wilkerson. 2021. Pandemic politics: Timing state-level social distancing responses to covid-19. *Journal of Health Politics, Policy and Law*, 46(2):211–233.

Nicolas Ajzenman, Tiago Cavalcanti, and Daniel Da Mata. 2020. More than words: Leaders' speech and risky behavior during a pandemic. *Available at SSRN 3582908*.

Hunt Allcott, Levi Boxell, Jacob Conway, Matthew Gentzkow, Michael Thaler, and David Yang. 2020. Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *Journal of Public Economics*, 191:104254.

Ravi Arunachalam and Sandipan Sarkar. 2013. The new eye of government: Citizen sentiment analysis in social media. In *Proceedings of the IJCNLP 2013 Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 23–28, Nagoya, Japan. Asian Federation of Natural Language Processing.

John M Barrios and Yael Hochberg. 2020. Risk perception through the lens of politics in the time of the covid-19 pandemic. Technical report, National Bureau of Economic Research.

Matthew A Baum. 2002. The constituent foundations of the rally-round-the-flag phenomenon. *International Studies Quarterly*, 46(2):263–298.

Donald J Berndt and James Clifford. 1994. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, USA:.

Gary Beverungen and Jugal Kalita. 2011. Evaluating methods for summarizing twitter posts. *Proceedings of the 5th AAAI ICWSM*.

Michela Bia and Alessandra Mattei. 2008. A stata package for the estimation of the dose-response function through adjustment for the generalized propensity score. *The Stata Journal*, 8(3):354–373.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Damien Bol, Marco Giani, André Blais, and Peter John Loewen. 2021. The effect of covid-19 lockdowns on political support: Some good news for democracy? *European Journal of Political Research*, 60(2):497–505.

Richard Brody. 1991. *Assessing the president: The media, elite opinion, and public support*. Stanford University Press.

Richard A Brody and Catherine R Shapiro. 1989. A reconsideration of the rally phenomenon in public opinion. *Political behavior annual*, 2:77–102.

Brandice Canes-Wrone, David W Brady, and John F Cogan. 2002. Out of step, out of office: Electoral accountability and house members' voting. *American Political Science Review*, pages 127–140.

Devin Caughey and Christopher Warshaw. 2018. Policy preferences and policy change: Dynamic responsiveness in the american states, 1936–2014.

US Census Bureau. 2012. *United States Summary, 2010: Population and housing unit counts*. US Department of Commerce, Economics and Statistics Administration, U.S. CENSUS BUREAU.

Koyel Chakraborty, Surbhi Bhatia, Siddhartha Bhattacharyya, Jan Platos, Rajib Bag, and Aboul Ella Hassanien. 2020. Sentiment analysis of COVID-19 tweets by deep learning classifiers - A study to show how popularity is affecting accuracy in social media. *Appl. Soft Comput.*, 97(Part):106754.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 6(2):e19273.

David DeFranza, Mike Lindow, Kevin Harrison, Arul Mishra, and Himanshu Mishra. 2020. Religion and reactance to covid-19 mitigation guidelines. *American Psychologist*.

Shana Kushner Gadarian, Sara Wallace Goodman, and Thomas B Pepinsky. 2021. Partisanship, health behavior, and policy attitudes in the early stages of the covid-19 pandemic. *Plos one*, 16(4):e0249596.

Oguzhan Gencoglu and Mathias Gruber. 2020. Causal modeling of twitter activity during covid-19. *Computation*, 8(4):85.

Guy Grossman, Soojong Kim, Jonah M Rexer, and Harsha Thirumurthy. 2020. Political partisanship influences behavioral responses to governors' recommendations for covid-19 prevention in the united states. *Proceedings of the National Academy of Sciences*, 117(39):24144–24153.

Pedro Henrique Calais Guerra, Wagner Meira Jr., and Claire Cardie. 2014. Sentiment analysis on evolving social streams: how self-report imbalances can help. In *Seventh ACM International Conference on Web Search and Data Mining, WSDM 2014, New York, NY, USA, February 24-28, 2014*, pages 443–452. ACM.

Keisuke Hirano and Guido W Imbens. 2004. The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164:73–84.

Paul W Holland. 1986. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Chhinder Kaur and Anand Sharma. 2020. Twitter sentiment analysis on coronavirus using textblob. Technical report, EasyChair.

Harleen Kaur, Shafqat Ul Ahsaan, Bhavya Alankar, and Victor Chang. 2021. A proposed sentiment analysis deep learning algorithm for analyzing covid-19 tweets. *Information Systems Frontiers*, pages 1–13.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Moritz UG Kraemer, Chia-Hung Yang, Bernardo Gutierrez, Chieh-Hsi Wu, Brennan Klein, David M Pigott, Louis Du Plessis, Nuno R Faria, Ruoran Li, William P Hanage, et al. 2020. The effect of human mobility and control measures on the covid-19 epidemic in china. *Science*, 368(6490):493–497.

Jeffrey R Lax and Justin H Phillips. 2009. Gay rights in the states: Public opinion and policy responsiveness. *American Political Science Review*, 103(3):367–386.

Jeffrey R Lax and Justin H Phillips. 2012. The democratic deficit in the states. *American Journal of Political Science*, 56(1):148–166.

Steven Loria. 2018. TextBlob documentation. *Release 0.15*, 2.

Cristian R Machuca, Cristian Gallardo, and Renato M Toasa. 2021. Twitter sentiment analysis on coronavirus: Machine learning approach. In *Journal of Physics: Conference Series*, volume 1828, page 012104. IOP Publishing.

Kamaran H Manguri, Rebaz N Ramadhan, and Pshko R Mohammed Amin. 2020. Twitter sentiment analysis on worldwide covid-19 outbreaks. *Kurdistan Journal of Applied Research*, pages 54–65.

Muvazima Mansoor, Kirthika Gurumurthy, Anantharam R. U, and V. R. Badri Prasad. 2020. Global sentiment analysis of COVID-19 tweets over time. *CoRR*, abs/2010.14234.

Micol Marchetti-Bowick and Nathanael Chambers. 2012. Learning for microblogs with distant supervision: Political forecasting with Twitter. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 603–612, Avignon, France. Association for Computational Linguistics.

Saif M. Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel D. Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Inf. Process. Manag.*, 51(4):480–499.

John E Mueller. 1970. Presidential popularity from truman to johnson. *The American Political Science Review*, 64(1):18–34.

John E Mueller. 1973. *War, presidents, and public opinion*. New York: Wiley.

Martin Müller, Marcel Salathé, and Per Egil Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse COVID-19 content on twitter. *CoRR*, abs/2005.07503.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Georgios Paltoglou and Mike Thelwall. 2012. Twitter, myspace, digg: Unsupervised sentiment analysis in social media. *ACM Trans. Intell. Syst. Technol.*, 3(4):66:1–66:19.

Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Ferran Pla and Lluís-F. Hurtado. 2014. Political tendency identification in Twitter using sentiment analysis techniques. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 183–192, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pages 451–463. The Association for Computer Linguistics.

Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Mrityunjay Singh, Amit Kumar Jakhar, and Shivam Pandey. 2021. Sentiment analysis on the impact of coronavirus in social life using the bert model. *Social Network Analysis and Mining*, 11(1):1–11.

James A Stimson, Michael B MacKuen, and Robert S Erikson. 1995. Dynamic representation. *American political science review*, pages 543–565.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2011. Sentiment in twitter events. *J. Assoc. Inf. Sci. Technol.*, 62(2):406–418.

Erik Tjong Kim Sang and Johan Bos. 2012. Predicting the 2011 Dutch senate election results with Twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 53–60, Avignon, France. Association for Computational Linguistics.

Tanmay Vijay, Ayan Chawla, Balan Dhanka, and Purnendu Karmakar. 2020. Sentiment analysis on covid-19 twitter data. In *2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, pages 1–7. IEEE.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

A Statistics of our Data

A.1 COVID Twitter Keywords

We list the COVID-related Twitter keywords and accounts tracked by Chen et al. (2020) in Table 8 and 9. They are used to retrieve the 1.01TB raw Twitter data.

Keywords used by Chen et al. (2020)	
14DayQuarantine	covidiot
CDC	epitwitter
COVD	flatten the curve
COVID__19	flattenthecurve
COVID-19	kung flu
China	lock down
Corona	lockdown
Coronavirus	outbreak
Coronials	pandemic
DontBeASpreader	pandemie
DuringMy14DayQuarantine	panic buy
Epidemic	panic buying
GetMePPE	panic shop
InMyQuarantineSurvivalKit	panic shopping
Koronavirus	panic-buy
Kungflu	panic-shop
N95	panicbuy
Ncov	panicbuying
PPEshortage	panicshop
Sinophobia	quarantinelife
Social Distancing	quarentinelife
SocialDistancing	saferathome
SocialDistancingNow	sars-cov-2
Wuhan	sflockdown
Wuhancoronavirus	sheltering in place
Wuhanlockdown	shelteringinplace
canceleverything	stay at home
china virus	stay home
chinavirus	stay home challenge
chinese virus	stay safe stay home
chinesevirus	stayathome
corona virus	stayhome
coronakindness	stayhomechallenge
coronapocalypse	staysafestayhome
covid	trump pandemic
covid-19	trumppandemic
covid19	wear a mask
covididiot	wearamask

Table 8: Keywords used by Chen et al. (2020) to track COVID-related tweets.

Accounts tracked by Chen et al. (2020)	
PneumoniaWuhan	WHO
CoronaVirusInfo	HHSGov
V2019N	NIAIDNews
CDCemergency	DrTedros
CDCgov	

Table 9: Accounts tracked by Chen et al. (2020) to retrieve COVID-related tweets.

A.2 Annotation Guidance for Policy Strictness

For each state, the annotators are asked to go to the official website that lists all COVID policies of the state. In most cases, the website lists all executive orders (EOs), proclamations, or other forms of policies issued during 2020 – 2021. Then the annotator is asked to read through the EOs that are related to COVID social distancing policies. For each relevant policy, the annotator is asked to record the start date on which the policy will take effect,[15] a brief intro of what kind of social distancing policy it is, and a real-valued score in the range of 0 (loosest) to 5 (strictest).

For the scoring criteria, we provide the following guides:

- Score 0: masks are optional, open the schools,, bars, gaming facilities, concert, and almost everything
- Score 1: State of emergency, limit gathering, close K-12
- Score 2: Open 50% capacity for retail business, open religious activities like churches to 50%
- Score 3: Open 25% capacity for retail businesses
- Score 4: Open only business for necessities such as supermarkets, only allow delivery and curbside services, gatherings have to be no more than 10 people
- Score 5: Strict stay at home policy, close every business

A.3 Accuracy of Twitter Sentiment Classifier

We list the detailed performance report of TextBlob and our COVID BERT in Table 10, including the overall accuracy, weighted and macro F1 scores, precision and recall for each class, and MSE of the average sentiment of random groups of 20 tweets. Note that since TextBlob predicts a real-valued number in the range of -1 to 1 for the sentiment, we regard [-1, -0.33) as negative, [-0.33, 0.33] as neutral, and (0.33, 1] as positive.

B Additional Analyses

B.1 Correlation across All Variables

We can see that, averaging over all 50 states, unemployment correlates the most with policy changes, which is consistent with our analysis in Section 5.1. Since different states may have different styles to

[15] For consistency, we record 0:01am of the first effective date, but not the 11:59pm of the previous day.

Model	Accuracy	F1 Score		Positive		Neutral		Negative		MSE on Groups
		Weighted	Macro	P	R	P	R	P	R	
TextBlob	23.35	16.67	19.70	20.34	10.62	20.67	85.19	74.07	6.45	0.43
COVID BERT	60.23	62.31	55.17	51.19	76.11	26.76	35.51	83.68	62.99	0.15

Table 10: The detailed performance report of the TextBlob baseline, and our COVID BERT model. We report the overall accuracy, weighted and macro F1 scores, precision (P) and recall (R) for each class, and MSE of the average sentiment of random groups of 20 tweets.

take sentiment into consideration when making policies, the effect of sentiment on policy changes over all 50 states is relatively mild.

For Twitter sentiment, it correlates largely with case numbers, and urbanization rate of the state.

Interestingly, the case numbers correlate with whether the state governor is a political ally of Trump.

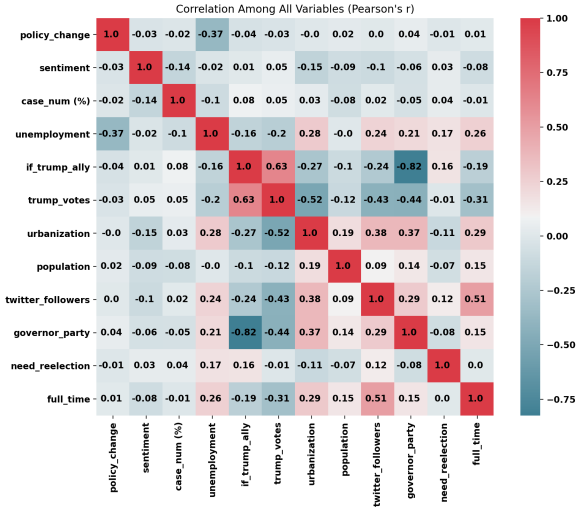


Figure 4: Correlation across all variables.

B.2 Alternative Causal Analysis Methods by Potential Outcomes Framework

There are two commonly used frameworks for causal inference, one is the do-calculus we introduced in Section 5.2, and the other is the potential outcomes framework (Rubin, 1974, 2005; Imbens and Rubin, 2015). We will introduce two alternative causal inference methods on our problem, using the potential outcomes framework.

Difference-in-Differences. One possible limitation of this study is that we treat the data in an i.i.d. way, following most existing studies. An improvement is to treat it as time series. For time series analyses, one commonly used method is the first-difference (FD) estimator, difference in differences (DID) (Abadie, 2005). Specifically, DID takes in the time series data of the cause X, effect Y, and confounders Z, and solves the following regression:

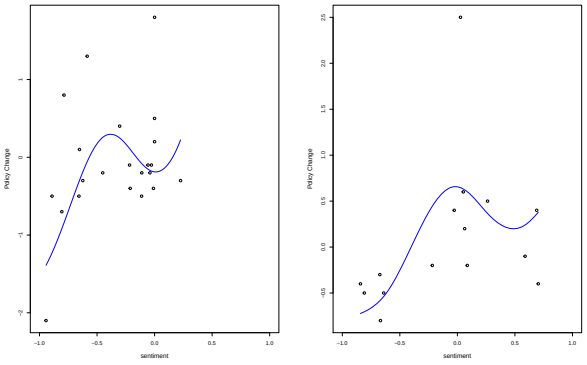
$$\Delta Y = \beta \cdot \Delta X + \Delta Z \quad (4)$$

$$Y_t - Y_{t-1} = \beta(X_t - X_{t-1}) + Z_t - Z_{t-1}\,, \quad (5)$$

where t is the time step, and β is the causal effect of X on Y.

After applying DID on all the policies, we obtain β scores for all states, and the top 5 states with largest β are Colorado ($\beta = 0.67$), Kentucky ($\beta = 0.23$), Wyoming ($\beta = 0.22$), Oregon ($\beta = 0.19$), North Carolina ($\beta = 0.17$), Michigan ($\beta = 0.14$), and New York ($\beta = 0.13$).

Continuous-Valued Propensity Score Matching Another commonly used alternative for causal inference is propensity score matching. However, the challenge in our study is that the cause is not categorical, but takes continuous values. To this end, we follow the extension of propensity score matching to continuous treatment (Hirano and Imbens, 2004; Bia and Mattei, 2008). We adopt the stata package of Bia and Mattei (2008) for continuous-valued propensity score matching. The resulting prediction of policies based on Twitter sentiment is a polynomial function with an order of three. As examples, We show the predictions of Texas (TX) and Michigan (MI) in Figure 5.



(a) Model of TX. (b) Model of MI.

Figure 5: Causal models by continuous-valued propensity score matching of TX and MI.