Appendix 1. Experimental Design

Table A-1. Co-occurrence matrix showing double-coding status among coders.

	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12
C01	1723	1723	554	555	0	0	0	0	0	0	0	0
C02	1723	1723	554	555	0	0	0	0	0	0	0	0
C03	554	554	1755	1752	0	0	0	0	0	0	0	0
C04	555	555	1752	1755	0	0	0	0	0	0	0	0
C05	0	0	0	0	2362	2362	800	800	0	0	0	0
C06	0	0	0	0	2362	2362	800	800	0	0	0	0
C07	0	0	0	0	800	800	2500	2500	0	0	0	0
C08	0	0	0	0	800	800	2500	2500	0	0	0	0
C09	0	0	0	0	0	0	0	0	2089	2087	796	796
C10	0	0	0	0	0	0	0	0	2087	2089	794	794
C11	0	0	0	0	0	0	0	0	796	794	2187	2187
C12	0	0	0	0	0	0	0	0	796	794	2187	2187

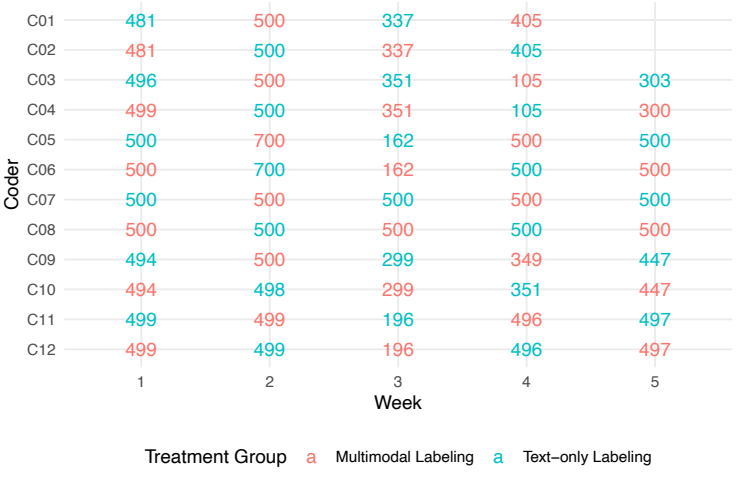

Figure A-1. Coders' Treatment Status Over Weeks. Twelve coders work on the task for five weeks. Each coder is assigned 500 posts to code per week. Coders alternative between the treatment and the control groups. The numbers in the figure show the actual number of posts coded by each coder-week.

Appendix 2. Additional Results

Table A-2. Evaluation Metrics (with 95% Credible Intervals from 1000 bootstrapping)

Metrics	Mean	5th Percentile	95th Percentile
ΔT	0.19	0.15	0.24
ΔT_v	0.14	0.10	0.19
ΔR	0.04	0.03	0.05
ΔI	0.09	0.00	0.19
ΔP	-0.03	-0.11	0.05

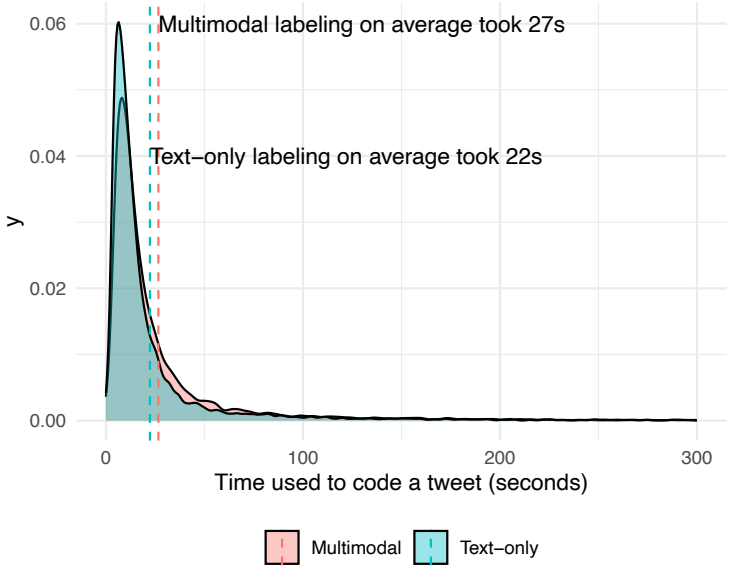

Figure A-2. Time taken for coders to label a tweet. On average it takes 5 more seconds to code a tweet with non-text features than text-only.

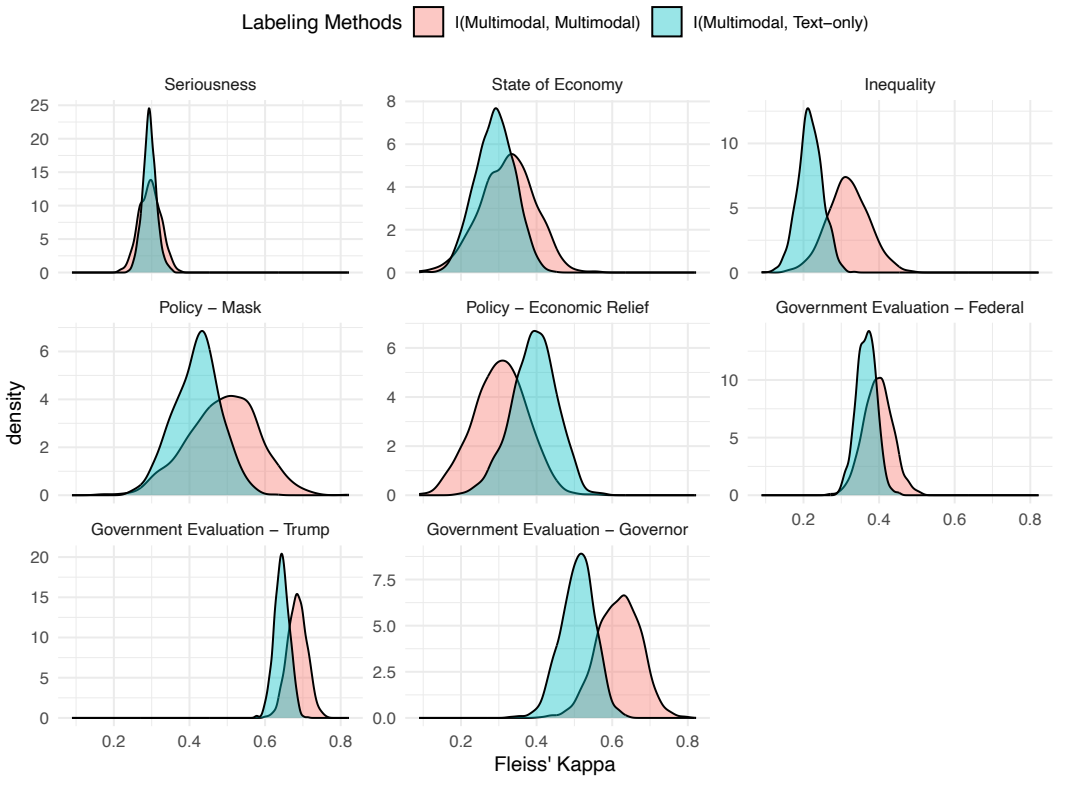

Figure A-3. Distribution of Fleiss' kappa values capturing intercoder reliability by label. Intercoder reliability is higher on every label with the exception of the label for whether the tweet is about an economic relief policy.

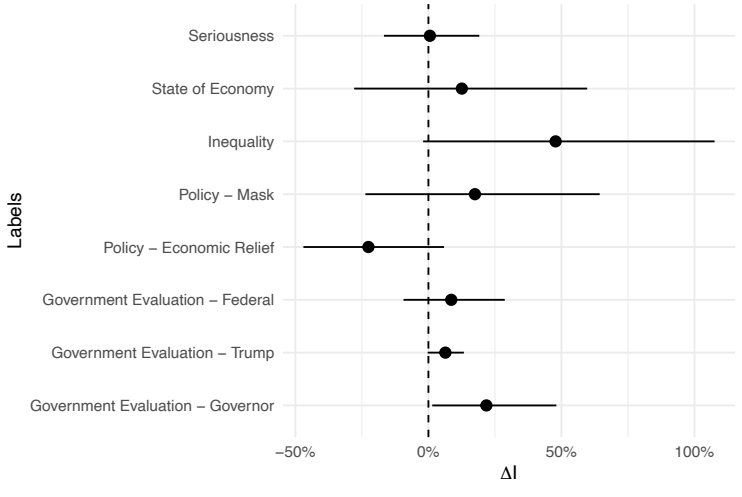

Figure A-4. Differences in intercoder reliability between multimodal and text-only conditions (ΔI, x-axis) by label (y-axis). Horizontal bars represent 95% confidence intervals based on bootstrapped standard errors sampled from two coders per tweet.

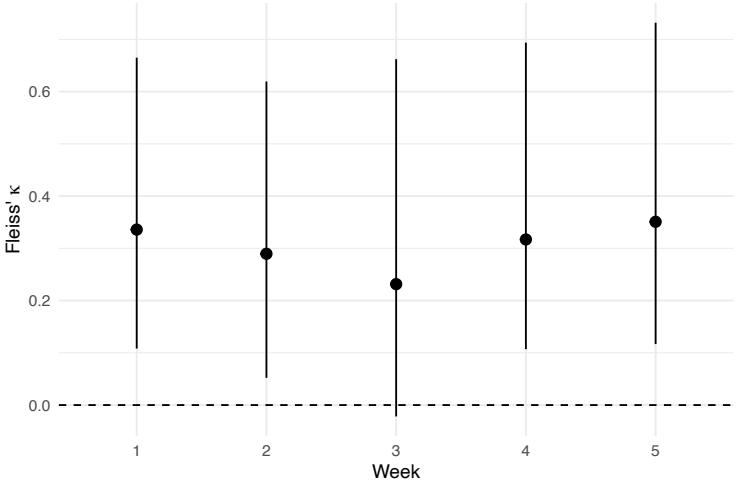

Figure A-5. We find no change in Intercoder Reliability over time. We examine whether coders tend to agree with one another more (regardless of treatment status) over time in our experiment. The figure shows no significnat change of Fleiss' κ (y-axis) from Weeks 1 to 5 (x-axis). The line ranges show the 95% Credible Intervals from 1000 bootstrapping.

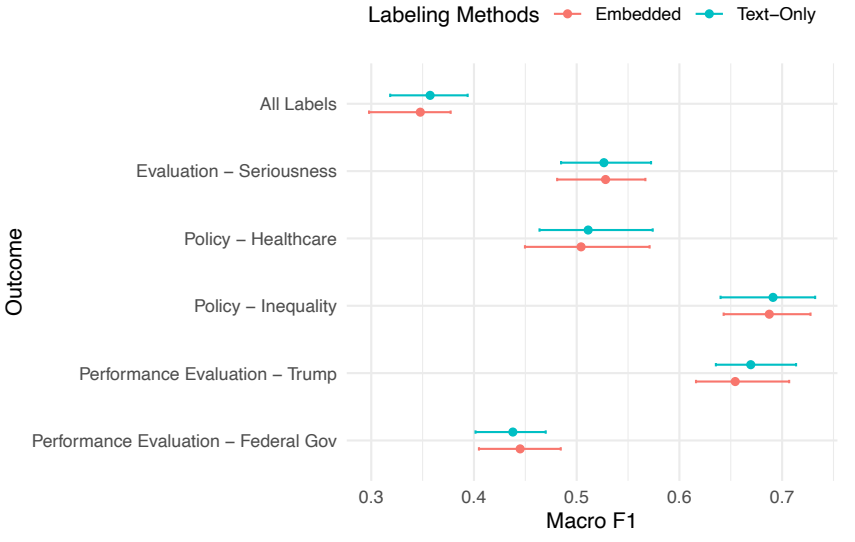

Figure A-6. Difference in text-based classifier performance (BERT algorithm) trained on text-only and multimodal data, and evaluated on hold-out sample of posts. Y-axis indicates either full multilabel classifier performance ("All Labels") or multilabel classifiers' performance trained with subsets of labels. Horizontal bars indicate two standard errors, calculated based on 100 cross validated splits.

Appendix 3. Label Frequencies

Table A-3. Number of tweets assigned to subset of labels (rows) by treatment condition (columns)

Label	Text-Only	Multimodal
Evaluation of COVID-19 seriousness		
Taking COVID-19 Seriously	1300	1201
Not taking COVID-19 Seriously	168	149
Concerns about the economic consequences of COVID-19		
State of economy	255	234
Inequality of pandemic	338	330
Attitudes towards COVID-19 policies		
Healthcare - Negative	252	241
Healthcare - Positive	107	107
Masks - Negative	21	18
Masks - Positive	96	99
Economic Relief - Negative	104	95
Economic Relief - Positive	79	98
Political support related to the handling of COVID-19		
Federal Government - Neutral	168	138
Federal Government - Negative	406	422
Federal Government - Positive	35	55
Trump - Neutral	218	241
Trump - Negative	704	740
Trump - Positive	130	143
Governor - Neutral	75	80
Governor - Negative	105	133
Governor - Positive	22	25
Valid Response		
Not enough information	1179	710
Number of quad-coded tweets	2351	2351
Number of double-coded tweets	7914	7914
Number of unique tweets	10,265	10,265
Total number of entries	12,616	12,616

If multimodal content can improve human annotation by empowering coders to better understand the content of a social media post, we might expect to see the reduction in posts annotated with the "not enough information" label, as we do in Table A-3 (1179 in the text-only condition, whereas there are only 710 in the multimodal condition). We further calculate the proportion of tweets that were labeled as "not enough information" in the text-only condition which were given a substantive label in the multimodal condition, broken out by whether both coders or just one indicated there was not enough information. These results, displayed in Table A-4, support the conclusion that the

additional of non-text content helps the coders better annotate the tweets.

Table A-4. Proportion of tweets that were classified as not enough information by coders in the text-only condition (rows) which were given a valid label in the multimodal condition (columns).

	Valid label in multimodal	
NEI in text-only	At least one coder	Both coders
At least one coder	89.96%	27.89%
Both coders	100%	81.25%

Looking in more detail at the association between labels chosen in the multimodal condition for tweets that had at least one coder indicate there was not enough information in the text-only condition, we find that the distribution of valid labels largely mirrors that found across the overall data. As illustrated in Figure A-7, the most commonly applied label pertains to tweets that took the pandemic seriously, followed by evaluations of Trump and the Federal Government.

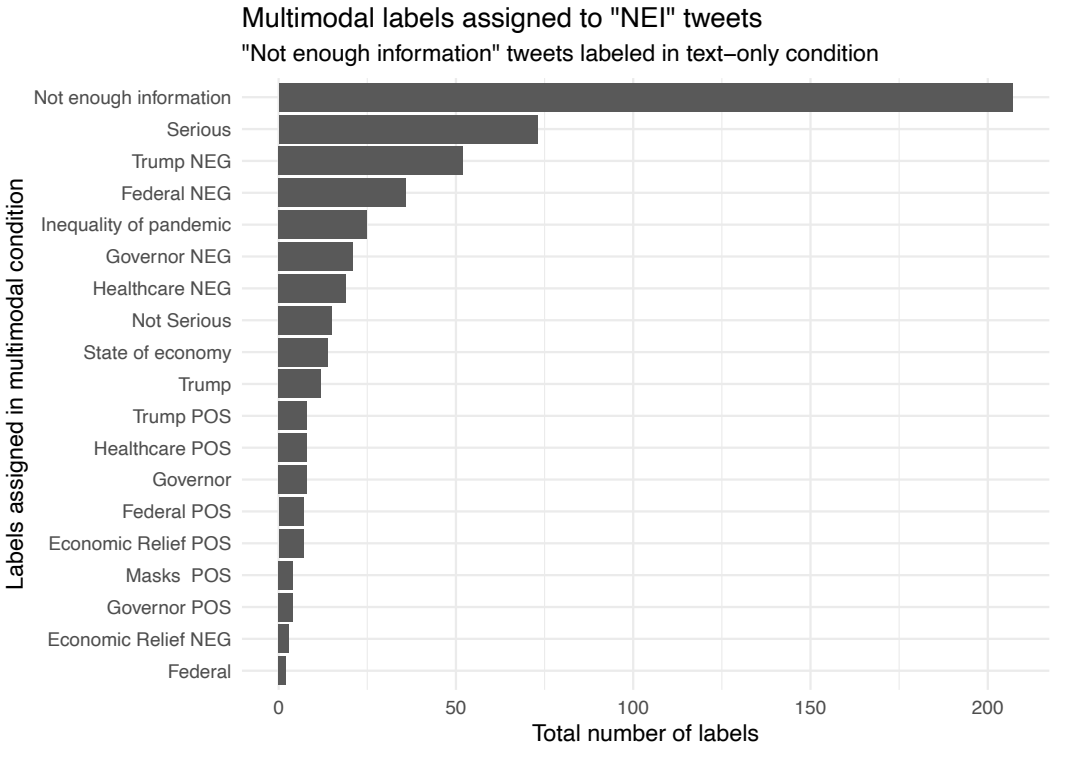

Figure A-7. Total count of labels assigned to tweets which at least one coder indicated were "not enough information" in the text-only condition.

Appendix 4. Full Description of the Annotation Task

We design a codebook, shown in Table A-5, that characterizes tweets as falling into six categories. The first two categories, "current situation" and "information" mainly measure *factual beliefs*. We evaluate whether a tweet discusses the current situation of the pandemic along different dimensions, and whether it contains factual information, misinformation, or conspiracy theories.

To measure *policy positions*, we evaluate whether a tweet shows support for, or objection to, a set of important policy issues such as a mask mandate and the closure of public spaces. To measure *political support*, we evaluate whether a tweet expresses approval or disapproval of the handling of the pandemic by a each of set of politicians and whether it expresses trust or distrust of a each of set of political and professional institutions. Finally, we include additional categories that evaluate whether a tweet discusses the influence of foreign entities or contains bias or hate speech in relation to Covid-19. Note that a tweet may be assigned multiple labels. For example, a tweet can simultaneously state a factual belief that the disease is not serious while also expressing approval of Trump's performance in addressing the pandemic.

Table A-5. Codebook overview

Category	Issue
Current situation	Taking the pandemic seriously or not
	Attitudes towards opening up/ closing down the economy
	Inequality of the pandemic
Information	Contains information, misinformation
	Promotes a conspiracy theory
Policy issues	Healthcare, masks, social distancing
	Closure of schools, churches, and public space
	Economic relief
Government performance	Election
	Evaluate the performance of:
	Federal government, Trump, governors, state or local policies
Biden	Mentions or expresses sentiment towards the presidential candidate
Institutional trust	Expresses trust or distrust of CDC, experts, WHO, and the media
Foreign entities	Mentions or expresses sentiment towards entities: China, Europe, Russia
Bias or hate speech	Express prejudice (or its rejection) towards Asian-Americans or immigrants

Appendix 5. Instruction to Coders

In this appendix, we reprint the instructions to coders.

Introduction

The current COVID-19 crisis provides the largest change in mass public behavior, and opinion, at the individual level the world has ever seen. In the United States, initial polls provided evidence of a wide partisan divide on opinions over the risk posed by the virus. But little is known about how public opinion got to this polarized point, and whether it was driven by consumption of different information, or by a difference across partisan groups in willingness to believe information from similar sources.

In this project, we will study how the public updates their opinions on the seriousness of COVID-19, as well as their opinions on the efficacy of restrictions on social and economic activity. And looking at polarization more broadly, we also examine their views of inequalities arising or made evident by the pandemic.

Task Description

We are asking for your help to code a set of tweets we think might be related to COVID-19. We are interested in labelling their relevance and sentiment on seven (non-mutually-exclusive) categories. Within each category there are usually several specific points we are interested in coding for.

1. Does the tweet contain an assessment of the seriousness of the ***current situation:*** *which includes comments on whether the tweeter wants to open or close the economy, and whether they express a view of the impact of COVID-19 on the state of the economy or on the inequality of the impact*?
2. Does the tweet mention specific ***policy issues*** *(such as civil liberties, access to healthcare, or the use of masks)*?
3. Does the tweet contain **factual information, misinformation, or a conspiracy theory**? 4. Does the tweet evaluate **government performance** as it relates to the crisis? This could be the performance of the federal government in general, or a specific governor, or the policy of a specific state.
4. Is the tweet about **Joe Biden**?

5. Does the tweet express a view about different **institutions** relevant to the COOVID-19 crisis (for example**,** the **CDC**)?
6. Does the tweet mention or express a sentiment towards **foreign actors** with respect to COVID-19 in the US?

Thus for each tweet, you could give the tweet anywhere from one label (e.g. 'irrelevant'), to many labels. A tweet could conceivably discuss or mention several of the seven categories above, and/or could accordingly receive multiple labels within any given category. The set of labels will be provided for you to choose from.

As we are only interested in tweets about the situation in the US, if the tweet is not about COVID-19 in the US or if you have not enough information to believe that it is, we would like you to indicate that in the **relevance** category and move on to the next tweet. Some tweets may be about COVID-19, but not be US-specific; and some tweets may simply not be about COVID-19.

In the online labeling app, we show you the text of the tweets along with embedded media (e.g., image, video, links to external web pages). We expect you to use all available information to make decisions on coding and indicate which of these pieces of information you used in the **methods** tag.

In some cases, you will see retweets of public officials, news outlets, or other accounts that are not owned by individuals. In these cases, you should consider the retweet an endorsement of the content being shared and score it accordingly. For example, if an individual retweets a post by the CDC providing guidance on how to socially distance, you should infer that the individual endorses this message and code the tweet accordingly.

In the following sections, we define each category of labels and provide examples.

Current Situation

This category is designed to capture the overall impression of the pandemic, ranging from the health risks to the impact on the economy. Labels include whether the author of a tweet takes the pandemic seriously, whether the author expresses a desire to reopen the economy or to maintain / extend social distancing policies, and two broad labels that capture statements about the impact of the virus on the economy writ large, or on inequality specifically. An example of a tweet indicating that the author takes the pandemic seriously is given below. Note that the author is speaking as a medical professional asking individuals to practice social distancing by not going to the ER if they have a

cough and a fever.

Figure A-8. Example tweet that takes COVID seriously

Note that a tweet can be assigned multiple labels. In the tweet below, this author takes the pandemic **seriously**, and is in favor of **waiting to re-open the economy**. While it may be tempting to assume that a tweet which is labeled as taking the pandemic seriously should also be in favor of waiting to open up, you should not assume this. We only want to label 'wait to open up' those tweets that explicitly suggest that. In the context of the below tweet, we can infer this advocacy by reading the article that the user asks others to read.

Figure A-9. Example tweet that favors waiting to re-open the economy

Finally, the tweet below is an example of one that we would characterize as **taking the pandemic seriously**, talking about the **state of the economy**, and in particular emphasizing the **inequality**

implications of the disease.

Policy Issues

The second broad category of labels is more specific and focuses on the policy response to the pandemic at the federal, state, and local level. The specific policy issues we would like to identify include:

- Gov intrudes civil liberties: Is the tweet critical of government restrictions on civil liberties?
- Healthcare: Does the tweet indicate that the author is satisfied or dissatisfied with the availability and quality of healthcare services in response to the pandemic?
- Masks: Does the tweet express a view on the importance of masks? Does the tweet suggest that the author thinks masks are unnecessary?
- Social Distancing: Does the tweet suggest approval or disapproval of social distancing? Social distancing can include explicit policies regarding how far apart people must stay from each other, or more general policies on which businesses are essential, when bars and restaurants can be opened, restrictions on non-essential consumption such as barbershops / spas / theaters, etc.
- School Closure: Does the tweets suggest approval or disapproval of closing schools (or, opening them if closed)
- Church Closure: Does the tweets suggest approval or disapproval of closing churches (or, opening them if closed)
- Public Space Closure: Does the tweet suggest approval or disapproval of closing public spaces (or, opening them if closed). Public spaces include beaches, playgrounds, and parks.
- Economic Relief: Does the tweet indicate that the author holds an opinion (positive or negative) about how the government is handling the economic relief in response to the pandemic? This can include things like rent freezes, stimulus checks, etc.
- Election: does the suggest **anything** about elections in relation to COVID-19 (delays, vote by mail, other)?

Information/ Misinformation/ Conspiracy

The third broad category focuses on the provision of information in the tweets. This can include factual information (i.e., sharing details about the scientific facts of the virus or the policy response),

mis-information, or conspiracy theories (e.g., that Bill Gates designed and intentionally spread the virus). Note that in some cases, determining whether a tweet contains information or mis-information may not be possible. As such, there is an option to label the tweet as "Information - unsure: Contains information relevant to the pandemic which you are not sure if it is true or false". Please do not code tweets as containing factual information if they are purely anecdotal, such as tweets that claim the user has the virus. An example of factual information is given below. Note that this is also an example of a tweet that takes the pandemic seriously, as discussed above.

Figure A-10. A tweet that shares **information**

An example of a tweet with questionable information is given below.

Govt Performance

The fourth broad category of labels pertains to how the user views the performance of different government agents in their response to the pandemic. The labels are divided into neutral, positive, or negative sentiments toward how individuals in the federal government (i.e., Senators or cabinet officials), Trump, governors, and local policies have responded to the pandemic. An example of a tweet containing negative sentiment toward both Trump and the federal government is given below.

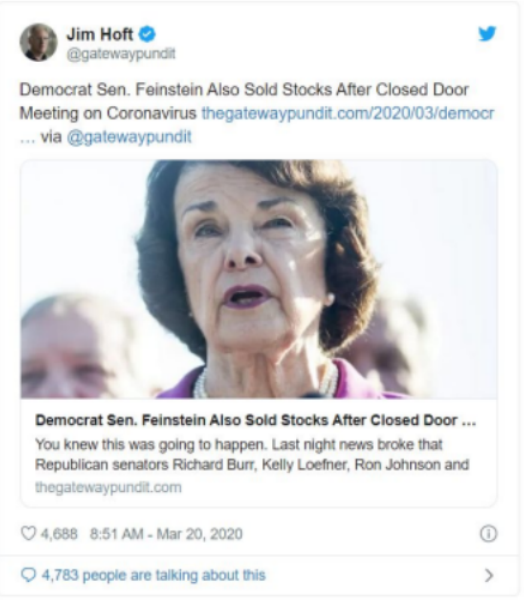

Figure A-11. A tweet that shares questionable information

Figure A-12. A tweet with negative sentiment towards Trump and the federal government

Biden

We are interested in a subset of tweets that pertain to the tweeter's assessment of Joe Biden. While Biden is not responsible for policy during the pandemic, we expect that users reference him specifically in the context of the 2020 presidential election, likely by talking about how he would have handled the situation. This tag is for tweets about Biden that are relevant to the pandemic - if it is just a general statement about Biden, or something about Biden's policy positions or actions unrelated to COVID-19, then the tweets would be irrelevant (or, at least NOT labelled as being about Biden).

Institutional Trust

The fifth broad category of labels is similar to the fourth, except that instead of pertaining to leadership's response to the pandemic, it pertains to non-leadership entities, including the CDC, the WHO, high-profile experts in general, and the media. These labels are intentionally broad, asking coders to identify tweets that contain either neutral, positive, or negative sentiments toward these groups. However, if the tweet does not refer to the pandemic at all, do not apply these labels. An example of a tweet expressing negative sentiment toward the WHO is given below.

Figure A-13. A tweet expressing negative sentiment toward the WHO

Foreign Entities

The sixth broad category of labels pertains to foreign actors with respect to COVID-19 *in the US*. As above, these labels should only be applied to tweets that mention a foreign entity in the context of the pandemic. Note that we are interested in opinions expressed by people in the US, offering opinions about foreign actors. This could be a person in the US suggesting that China should have been more transparent about the virus, or blaming travelers from another country for bringing the virus into the US.

Note that we are **not** interested in tweets that appear to be written by non-US based users. We are only interested in those that are from a US-based user talking about a foreign country, *as that country relates to the pandemic in the US*. The example of a tweet expressing negative sentiment toward

the WHO (above) is also a tweet containing negative sentiments toward China.

A tweet mentioning only a foreign entity (without referring to the COVID-19 situation in the US) should be labeled *irrelevant* (see description of the "Relevance" category below) unless it meets both of the following two criteria: (1) there's is no evidence *based solely on the tweet* that the tweet is written by a user located outside the US. (2) it implies actions in or by foreign countries or actors influence the COVID-19 situation in the US.

An example of a tweet that we are **NOT** interested in is given below. While the tweet is about COVID-19 and a foreign actor, it is not written by someone living in the United States, nor does it say anything about that foreign actor affecting the US.

Figure A-14. A tweet about foreign entities not related to the US

An example of a tweet that we are interested in is given below. While the tweet only discusses the COVID-19 situation in China and does not explicitly mention that in the US, it is considered expressing a negative sentiment about a foreign entity, China, because it suggests China's cover-up of COVID-19 severity which has implications for the COVID-19 situation in the US.

Figure A-15. A tweet expressing a negative sentiment about a foreign entity

Bias or Hate Speech

If the tweet attempts to blame in any way the pandemic on either Asian/Asian-Americans or immigrants, or uses the pandemic as justification for expressing a negative view about a group, it should be coded as negative sentiment toward these groups. If the tweet defends these groups *in the context of the COVID-19 pandemic*, it should be coded as positive sentiment toward these groups.

Relevance

The set of labels described above are meant to be reasonably exhaustive. However, there are many tweets that will not fit into these categories. These may include tweets that do not include enough information, are related to COVID-19 in a dimension that the above categories don't capture, or are simply irrelevant. We are NOT interested in tweets that are about life in general during the pandemic. Please code these as irrelevant. An example of such a tweet is given below. Note that while this tweet is about COVID-19, and appears to be set in the United States, it is simply making a joke about life during a pandemic.

Method

The labels described above are designed to capture the substantive content and perspective of the Twitter user who is tweeting about COVID-19. The "Method" category instead is interested in how

Figure A-16. An irrelevant tweet

you, the coder, made your determination. There are three options: "image", "video", and "followed a link". In all cases, we expect that you make your determination first and foremost by relying on the text of the tweet itself. However, if you use an embedded image, video, or link when making your determination, please indicate as such with this category.

Unsure

Finally there is a checkbox labeled "Unsure". This is not meant to be its own label for a given tweet. Rather, you should do your best to label each tweet according to the guidance provided above and in the codebook. After making your selection(s), if you feel unsure about the tweet you may click this checkbox.

Using the App

We have developed an online coding app to help you in your task. You will be given a unique username and password that you use to log into your account. This allows you to pause the work and return to it as needed. Each tweet you code is automatically saved (please ensure you have a reliable internet connection).

When you first log in, you will see a greeting page that contains information on your coding progress, as well as a chart displaying when you have been working on this task.

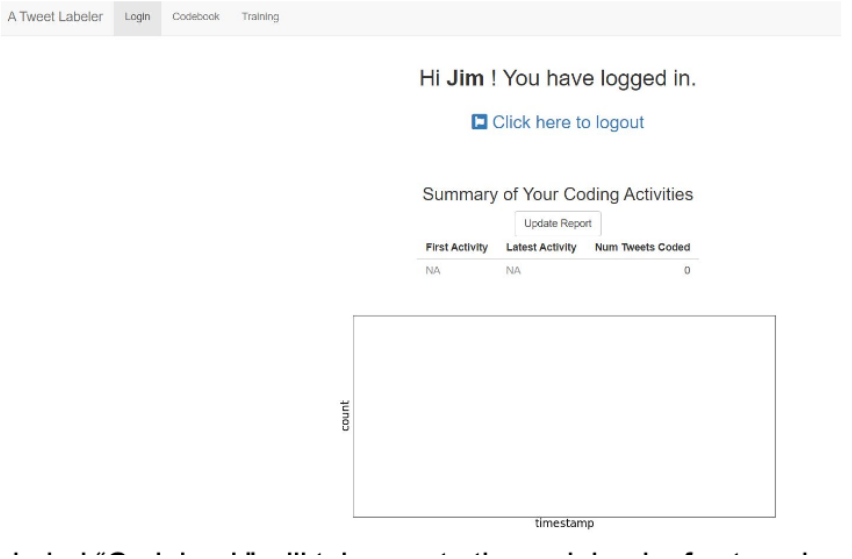

Figure A-17. Login page of the labelling app

The second tab labeled "Codebook" will take you to the codebook of categories that you should refer to with questions about the different labels. You may either refer back to this tab as needed, or you can export the code book to softwares of your choosing (such as Excel or PDF) or print out a physical copy.

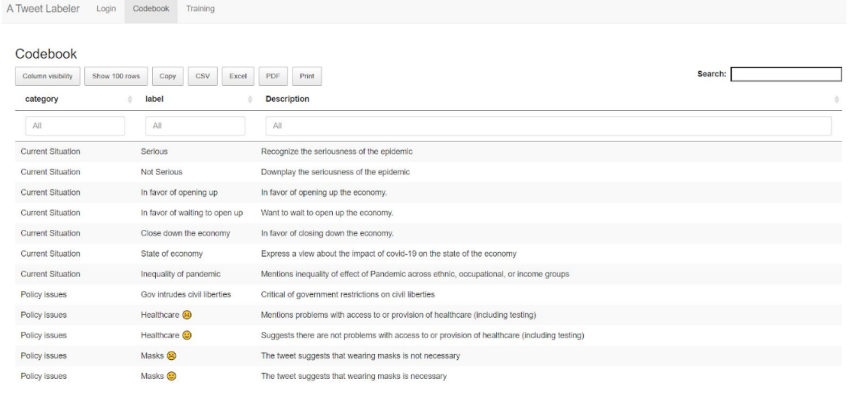

Figure A-18. Codebook page in the labelling app

The third tab is the coding interface and is comprised of two columns, as highlighted in the picture below. Column 1 contains the tweets themselves, along with any links contained therein.

Column 2 contains dropdown multiple selection boxes for the categories listed above. Note that you can select multiple labels both across categories, as well as within a given category (i.e., you can code a tweet as expressing negative sentiment about Trump, the federal government, and China altogether).

Figure A-19. Coding panel of the labelling app

In some cases, tweets will not be embedded, and will show up as raw text instead (see example below). Code these as well and, if necessary click on any provided links to aid in making your determination.

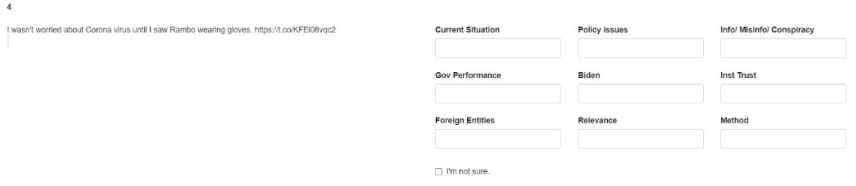

Figure A-20. Example case when the embedded tweet fails to load

You can select how many tweets to view per page at the top of the page, and can navigate freely. Please make sure to code all tweets to the best of your ability. Each time you select a label, it is automatically saved. However, if you made a mistake or changed your mind, you can adjust your label and it will also be saved. If you need to take a break, you can log out and be confident that when you log back in, all your progress has been saved.