Political Science Research and Methods (xxxx), **xx**, 1–34

CAMBRIDGE
UNIVERSITY PRESS

RESEARCH NOTE

Labeling Social Media Posts: Does Showing Coders Media Content Produce Better Human Annotation, and a Better Machine Classifier?

Haohan Chen,[*,1,2] James Bisbee,[3,2] Joshua Tucker,[2,4] and Jonathan Nagler[2,4]

[1]Department of Politics and Public Administration, The University of Hong Kong
[2]Center for Social Media and Politics, New York University
[3]Department of Political Science, Vanderbilt University
[4]Wilf Family Department of Politics, New York University
[*]Corresponding author. Email: haohan@hku.hk

Abstract

We investigate how to improve human annotation of social media posts. We focus on the effect of giving coders access to posts' media content on the quality of labeling. We built a web application for labeling posts and randomly assigned coders to a treatment group that had access to media content, such as embedded headlines and images in posts, and a control group where only the raw text of posts was rendered. Our results show that while access to the media content slightly improved intercoder reliability, it did *not* improve the performance of text-based machine learning classifiers trained with such labeled data. Our results suggest caution in providing features to coders that will not be available to text-based classifiers.

Keywords: Text and Content Analysis; Measurement; Mass Media and Political Communication

Social media data are extensively used as measures of public opinion and elite messaging in computational social science research. To utilize these data, a crucial task is labeling posts for concepts of interest. As the size of such data are generally larger than what researchers can exhaustively label manually, a machine-assisted supervised learning approach is commonly adopted: human coders label only a small subset of the data, then a machine classifier is developed to label the rest of the corpus. Typically, such an approach takes the following steps:

1. A large sample of posts of interest are gathered.
2. The researcher develops a codebook describing the labels to be applied to these posts.
3. Human coders manually annotate a small subset of the sample according to the codebook.
4. Supervised machine learning classifiers are trained using the coders' annotated subset.

5. The best machine classifier is used to label the remainder of the posts.

To minimize measurement error, researchers need to pay close attention to each step of the workflow. There have been a variety of methodological innovations proposed for improving different components of the workflow. For example, in Step 1, keyword expansion methods have been developed and applied to generate useful samples when trying to find the corpus of posts about a specific topic (King, Lam, and Roberts 2017). For Steps 4 and 5, there is constant innovation in the machine learning algorithms that are used to train machine classifiers (Devlin et al. 2018; Liu et al. 2019; Wu et al. 2022). However, little attention is paid to Step 3: there has been a lack of systematic evaluation on how to improve the quality of the human annotations that constitute the training data. We believe this gap is important for the simple reason of "garbage in, garbage out." If the human annotated training data is noisy or biased, then one should not expect the classifier to perform well, meaning that the resulting variables will be measured with greater error, and downstream analyses will be affected.

Our paper attempts to improve upon Step 3 of this workflow. Specifically, we focus on the costs and benefits of showing content associated with posts (e.g., headlines, images) beyond simply the raw text of the post to human coders. With only a few exceptions (Miller, Linder, and Mebane 2020), the majority of existing work applies machine learning classifiers to training data annotated based solely on the raw text of the post. However, this arguably misses out on a non-trivial part of the information contained in each post. A considerable number of posts also contain additional media content (i.e., images, videos, links) that will support, supplement, or augment the text message.[1] In some cases, this additional content can provide context information that can significantly change one's interpretation of a post's raw text. As a result, it is reasonable to hypothesize that including media content in the data labeling process can benefit researchers by generating more accurate labeling of the training and test datasets. In fact, as common post labeling tasks involve subjective perception of the meaning of the post, we will ultimately treat posts labeled with all information contained in the post as closer to 'truth.' However, what we examine in this paper is which method produces a training dataset that will lead to a classifier best able to correctly classify posts. We return to this in our section on results.

Including media content is costly: infrastructure needs to be set up to show coders posts with the

1. We use the term 'media content' to refer to images, videos, and snapshots of linked content.

media content intact (as opposed to cheaply placing the post's raw text in a spreadsheet). Furthermore, coders need to spend time and energy going through the media content, which can affect their efficiency and job satisfaction. Finally, if the machine learning algorithm is solely text-based, it is possible that media-enriched labeled training data might do more harm than good by labeling posts based on content to which the algorithm does not have access. If a human labels a posts based on media content, the classifier is being fed 'error' as the label assigned does not match the text the classifier is trying to relate to the label. Hence, there are two important question to ask is. First, does including media content produce a labeled dataset that is of higher quality than one produced without media content as measured by inter-coder reliability? High levels of intercoder reliability suggest that humans are being given enough information to come to the same coding decision. Second, does including media content produce a labeled dataset that, when used as a training dataset, generate a better (text-based) classifier than a classifier trained on data that is produced only by labeling based on text?

We attempt to answer these questions by experimentally evaluating the benefits and costs of including media content in tweet labeling. In a recent tweet labeling task, we developed a web-based data labeling platform that allowed us to randomly assign tweets with media content intact to one set of coders, while the other set of coders had access to only the raw text. We then evaluate the performance of human annotation on several dimensions. Our primary purpose is to evaluate the downstream task of interest – the performance of the machine classifier trained on the media-enriched versus raw text-only labeled data. But we also evaluate the two sets of labels for inter-coder reliability, the proportion of data that coders were not able to classify (i.e., labeled as "could not tell"); the time spent on each task; and self-reported job satisfaction for the human coders.

Our findings suggests that giving coders access to posts' media content has both benefits and costs on the quality of human annotation, but that it harms the performance of machine learning classifiers. Specifically, the results of our experiment show that when coders have access to tweets' media content, there is a slight improvement in inter-coder reliability, a significantly lower proportion of missing data, and a slight improvement in coders' job satisfaction, though these benefits come at the cost of additional time being needed to complete the labeling tasks when media content is included (and therefore higher costs to researchers if labeling labor is being compensated at an hourly rate). However, when tweets coded with media content are used to train machine learning classifiers, they

return models with slightly worse performance. We believe that these conclusions generalize beyond our specific application to Twitter data, and are of broad interest to any applied research engaged in labeling online text that is accompanied by additional media.

1. Research Design

The Tweet Labeling Task

We experimentally investigate the effect of access to media content with a fairly complicated tweet labeling task conducted in the summer and fall of 2020. We recruited 12 coders to label 11,637 tweets related to COVID-19 in order to capture the tweet author's attitudes about the pandemic. Coders were required to evaluate whether the tweet contained content related to any of eight broad categories. And within each category, coders had multiple non-mutually exclusive labels that could be assigned to the tweet. Across all categories there were a total of 79 non-exclusive labels that could be assigned to a tweets.[2] Among them, we were interested 19 specific labels from our general categories of labels:

- **Current Situation: Evaluation of COVID-19 seriousness**: Coders examined whether authors of tweets took COVID-19 in the United States *seriously* or *not seriously*. In cases that authors do not express explicit evaluation of COVID-19 seriousness (which constitutes the majority of cases in the labeled dataset), neither of the above two labels apply ("not evaluating the seriousness of COVID-19").
- **Policy issues related to COVID-19**: Coders evaluated whether tweets express opinions (*approval* or *disapproval*) about the following policy issues: Healthcare policies, mask-wearing policies, and economic relief policies.
- **Evaluation of government performance** (*approval* or *disapproval*): Coders evaluated how different political entities handled COVID-19, including: The federal government, President Trump, governors, and state or local policies.
- **Current Situation: Evaluation of Economy and Inequality**: Coders label whether tweets express views about wanting to *open up* or *close down* the economy, and whether they mention *inequality* caused by the pandemic.

2. Figure 1 shows how coders can apply non-mutually exclusive labels to tweets.

Our Tweet Labeling Infrastructure

We designed a web application that would display the full tweet to coders – text and media content – for the tweet labeling task. To complete their tasks, coders needed to log onto our app using a web browser. Figure 1 shows how a tweet to be labeled is displayed to a coder in the app under the embedded tweet treatment condition. The tweet content and labels to be applied are displayed side-by-side. The tweet's text content is shown on the top-left and the possible labels are shown on the right (with drop-down menus). Importantly, for coders in the treatment condition, an embedded tweet is displayed following the tweet text on the left panel (using Twitter's oEmbed API).[3]

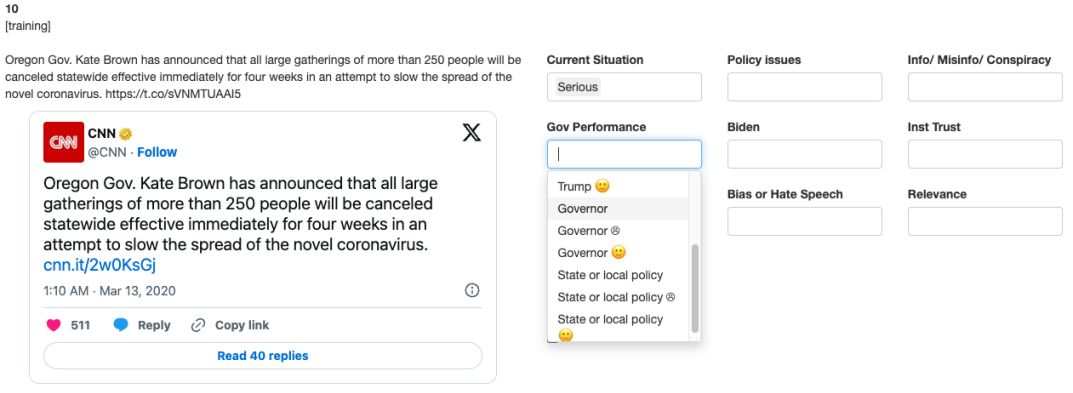

Figure 1. Interface of Our Tweet Labeling Infrastructure

The Experimental Setup

We split the labeling tasks into 5 assignments, given to our coders at the rate of one assignment per week. In each assignment, we presented half of the coders with only the text of the tweets (referred to as the "text-only" or the "control group" hereafter) and the other half with text along with embedded tweets (referred to as the "embedded tweet" or the "treatment group" hereafter). Figure 1 displays the tweet as it was seen by the treatment group. Each coder alternated week-by-week between labeling as part of the "text-only" control group and labeling as part of the "embedded tweets" treatment group.

In each assignment, coders were assigned 500 tweets, 200 of which were also assigned to 3 other coders and 300 of which were assigned to 1 other coder. Coders stayed in the same group throughout

3. Note that in rare cases where the tweet has been deleted or made protected, the oEmbed API will fail to render the tweet. These instances are rare in our experiment, and results are robust to dropping them or using them as control units.

the task so that each coder's assignment overlapped with the same 3 other coders throughout all 5 assignments.[4] Table 1 displays the co-occurrence matrix showing the total number of tweets each coder co-labeled with others.

Table 1. Assignment Overlap between Coders

	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12
C01	2000	1818	649	555	0	0	0	0	0	0	0	0
C02	1818	2500	999	905	0	0	0	0	0	0	0	0
C03	649	999	2507	2102	0	0	0	0	0	0	0	0
C04	555	905	2102	2518	0	0	0	0	0	0	0	0
C05	0	0	0	0	2501	2362	800	800	0	0	0	0
C06	0	0	0	0	2362	2500	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	2500	2288	1000	797
C10	0	0	0	0	0	0	0	0	2288	2448	995	795
C11	0	0	0	0	0	0	0	0	1000	995	2500	2188
C12	0	0	0	0	0	0	0	0	797	795	2188	2482

Figure 2 shows the treatment status of coders, and the number of tweets labeled, throughout the five weeks of our labeling task.[5] As shown by the number of tweets coders completed per week, we see that coders in general complied with the design, although some failed to finish a small proportion of their tasks.[6]

We assigned 12,000 tweets to coders and received 11,637 valid responses for our analysis. These tweets are a random sample of COVID-19-related tweets filtered by keywords. On average, each tweet is labeled by 2.5 coders. The labeled tweets can be divided into two datasets. First, 2351 tweets were labeled by exactly 2 coders from the treatment (embedded) group and 2 from the control (text-only) group. Second, 7914 tweets were labeled by exactly 1 coder in the treatment (embedded) group and 1 coder from the control (text-only) group.

We evaluate the collected data to answer two questions. First, does giving coders access to tweets' media content improve human annotation? Second, does it improve the machine classifier trained with the labeled data? We employ four indicators to capture the quality of human annotation and

4. We asked coders to complete their tasks independently and, to our knowledge, coders did not collaborate to finish the tasks.

5. We assigned the same fixed number of tweets to each coder. Coders may fail to complete the assignment.

6. Week 2 saw a few coders coding more than 500 tweets. That is because there was a delay in switching treatment groups. As a results, coders who started to work early for the week were placed in the wrong group. We drop these observations from our data analysis.

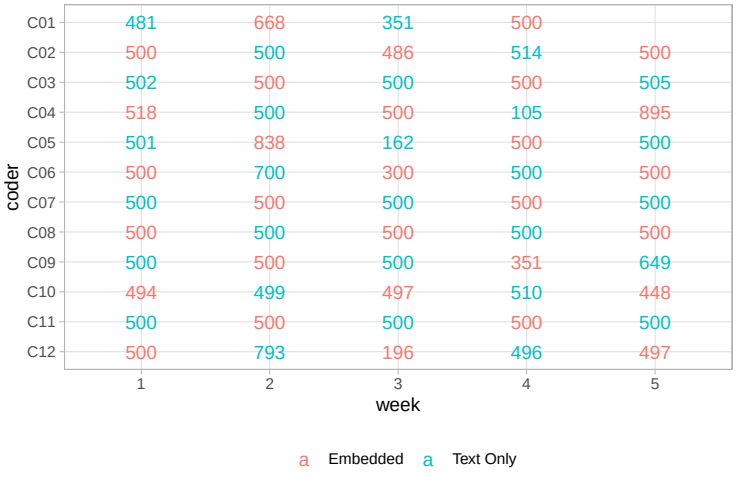

Figure 2. Coders' Treatment Status Over Weeks. Twelve coders work on the task for five weeks. Each coder is assigned 500 posts to code per week. Coders alternative between the treatment and the control groups. The numbers in the figure show the actual number of posts coded by each coder-week.

experience of the coders, and one for the quality of the machine classifier, as specified below:

- **Inter-coder reliability:** We compare inter-coder agreement within the treated and the control groups and between the two groups respectively and assess whether non-text features improve or erode agreement. We interpret higher levels of intercoder reliability as a measure of labeling quality.
- **Missing data:** We offer coders the option to indicate that they do not have enough information to label a tweet. We assess whether showing non-text features reduces the proportion of tweets that coders are unable to label.
- **Speed:** We compare the time it takes for coders to label a tweet in the treatment and control groups to understand whether showing non-text features affects the speed of task completion.
- **Coders' Job Satisfaction:** Over the duration of the experiment, we surveyed coders about their subjective experiences completing the task each week. We analyze whether being in the treatment or control condition influences how they rate their experience.
- **Performance of machine classifier:** We fit machine learning classifiers trained on tweets labeled in the treatment and control conditions separately, and test whether treatment status meaningfully influences classifier performance. Specifically, we use the Macro F1 score to measure model performance.

2. Results

Does Giving Coders Access to Media Content Improve Human Annotation?

Media content modestly improves inter-coder reliability. We start our evaluation by comparing the inter-coder reliability of labeled tweets by coders' treatment status (i.e., whether they have access to tweets' media content through embedded tweets in the application). Figure 3 compares the Fleiss κ of the two sets of labeled tweets. Higher values suggest that labels from the ebmedded treatment group had higher inter-coder reliability than labels from the text-only control group. Overall, access to media content sees a slight improvement in 6 out of the 9 sub-tasks. However, the 95% credible intervals of the differences obtained from bootstrapping all crossed zero.[7] We consider this a minor improvement that suggests that access to media content can get coders to agree with one another more.

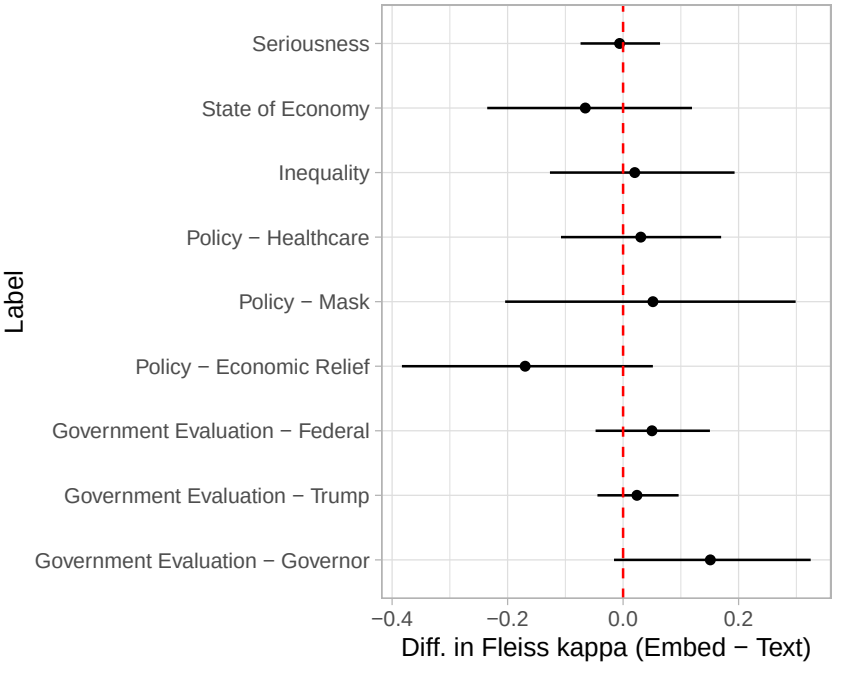

Figure 3. Comparing Fleiss' Kappas between tweets coded with and without media content. Accessing media content improves 6 out of 9 sub-tasks. But the bootstrapped 95% Credible Intervals of the differences all crossed zero.

7. To obtain boostratpped CIs, we performed sampling with replacement for 1000 times to generate 1000 bootstrapped datasets. For each dataset, we calculate the Fleiss's Kappas for the treated and the control groups respectively and took the differences. The estimated differences (dots in the figure) are the median of the bootstrapped differences.

Access to media content leads to a significant reduction of missing data. Does including media content provide additional information that helps coder assign at least one meaningful label to tweets? Our findings suggest that it does. In our labeling task, we instructed coders to select "not enough information" if, with reasonable effort, they were still unable to determine what labels apply to a tweet. Below are a few examples of tweets at at least one coder thought had insufficient information to label:

- *Not stopped. Not closed down. Not going to zero. [Link]*
- *A must read. Very diligent and data driven analysis showing fatality rates for prepared and unprepared countries [Link]*
- *That's actually a great idea. I'm going to do that too.*

Tweets that are ultimately labeled as "not enough information" should be considered missing data – this is the portion of labeled data that cannot be used to train machine classifiers since no meaningful label is applied. Figure 4 shows how the proportion of tweets labeled as "not enough information" changes when tweets are embedded (coders are given access to media content) and when only tweet text. For all but one coder (who never labeled a tweet as "not enough information" in either treatment condition), access to media content reduced the possibility of generating missing labels. And the increase in labeled tweets ranged from 1 percentage point to 6 percentage points.

Media content increases the time it takes to code. Another measure for evaluating coder performance differences across the embedded and text-only conditions is the time taken to label a tweet. One might expect that the embedded condition takes less time if the richer information environment speeds up the process of interpreting the substantive content of a tweet. If nothing else, a picture is worth a thousand words. But conversely, given that fewer tweets were labeled as "not enough information" in the embedded condition, we might instead expect that it takes coders longer since they are able to label more tweets overall. Indeed, this latter expectation is borne out in the data, with the average time taken to code a tweet in the embedded condition being two seconds longer than the average time in the text-only condition. As shown in Figure 5, the average time it takes to code under the text-only condition is 15 seconds, while that for the embedded condition is 17 seconds.[8] The obvious implication is that it might cost more to label tweets if media is included,

8. The increased time is statistically significant under a paired two-sample t-test.

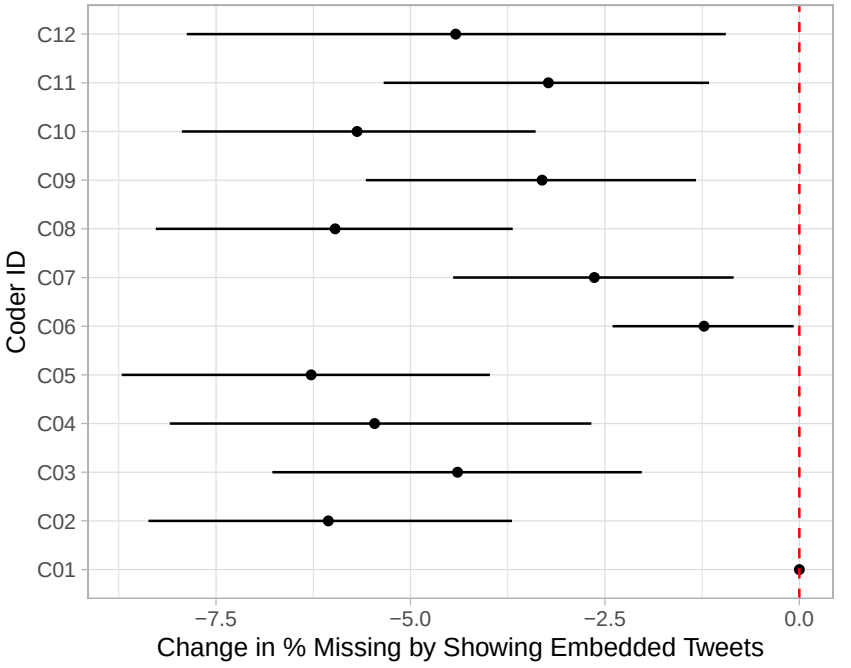

Figure 4. Embedded tweets reduce the proportion of missing label for all but one coder. Coder 1 (C01) didn't label any tweets as "not enough information" in either treatment condition. Hence, his estimated changes are 0 with no variance. Horizontal bars represent 95% Credible Intervals obtained through bootstrapping 1000 times.

though this is somewhat alleviated by the higher proportion of valid labels returned.[9]

Access to media content slightly improves coders' job satisfaction. Over the course of the experiment, we had weekly check-ins with our coders, during which we asked them to complete a brief survey. The survey asked coders to indicate their agreement with the statement that the past week's coding task was either "boring" or "inconvenient" on a scale from 1 to 5. Figure 6 shows coders' subjective perceptions of how boring or inconvenient the tasks were across the two groups. Coders found text only tasks slightly more boring and slightly more inconvenient. But the differences were negligible and nowhere near conventional levels of statistical significance.

In summary, judging by the quality of human annotation, giving coders access to tweets' media content provides some benefits, including a slight improvement in inter-coder reliability, and a significant increase in the number of valid data points. However, it also comes with costs, somewhat increasing the average time required to label a tweet.

9. The time per non-missing label returned is 16.5 seconds for text, 17.9 seconds for embedded.

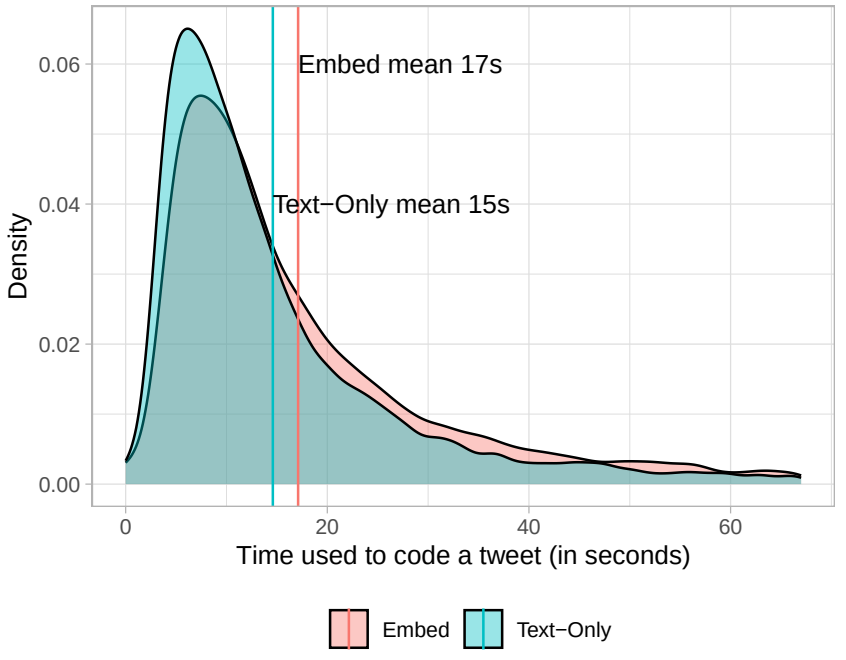

Figure 5. Time taken for coders to label a tweet. On average it takes 2 more seconds to code a tweet with non-text features than text-only.

Does Coders' Access to Media Content Improve Machine Classification?

Finally, we evaluate the intended output of tweet labels: the performance of the supervised machine learning models trained with the labeled data. For most studies, it is of researchers' primary interest to produce a reliable machine classifier and use it to label a larger set of posts that are not manually labeled. Do tweets labeled with media content available to the coder help train a better classifier than tweets labeled with only text available? Our experiment suggests the opposite.

We train two sets of machine classifiers using the labeled data our coders produced under the two experimental conditions. The classifiers are all built upon the Transformer model and, specifically, use the RoBERTa pre-trained model (Liu et al. 2019). We fit two types of models: (1) multi-label classifiers that include all labels of interest as non-mutually exclusive binary outcomes (namely the "full model") and (2) multi-class classifiers that separately predict a select set of mutually exclusive outcomes. The first row of Figure 7 gives the performance on the multi-label classifier; and the other five rows give the performance of the mulit-class classifiers.[10] To account for uncertainty in

10. We did not fit multi-class classifiers that separately predict *state of economy*, *policy (mask)*, *policy (economic relief)*, and *government evaluation (governors)* because of their extreme class imbalance (i.e., too few cases).

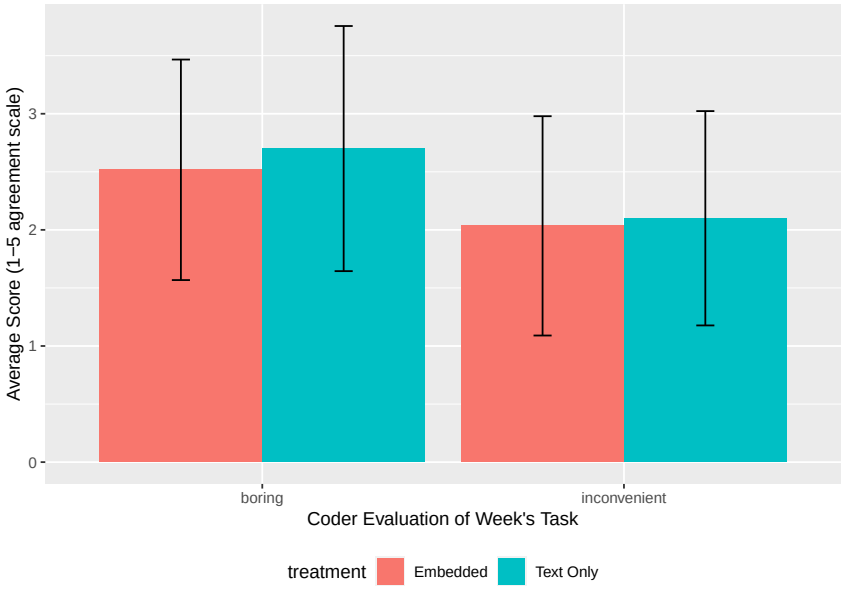

Figure 6. Average agreement with the statement that the coding task was "boring" (left) or "inconvenient" (right) among respondents in the embedded tweet treatment condition (orange) or text-only treatment condition (teal).

the data-generating process, for each outcome variable, we use 100 different random seeds to split the training and validation data. This gives us a credible interval instead of a point estimate for the models' performance metrics. We use Macro F1 scores as our evaluation metrics of interest, given class imbalance for all the tasks.

Figure 7 summarizes the results. Overall, tweets labeled with coders' access to media content do not train better classifiers. To the contrary, the average Macro F1 scores of media-available (i.e., embedded) tweets are lower than those of text-only tweets in 4 out of 5 tasks (including the "full task" where we predict all the labels with a multi-label classifier).

It might seem puzzling why additional information to coders makes the classifier worse. While our experiment is unable to answer questions about the mechanism, we speculate the reason to be the "information gap" between coders and the machine: in the situation where coders have access to media content of tweets but the machine classifier does not, the machine learning classifier might fail to put weight on appropriate features, or might put excessive weight on inappropriate features.

Given our research design, we want to emphasize that all we can infer from the results is that giving coders access to media content does not make better machine classifiers *when the machine learning algorithm cannot learn from the same media information*. This conclusion might change if we

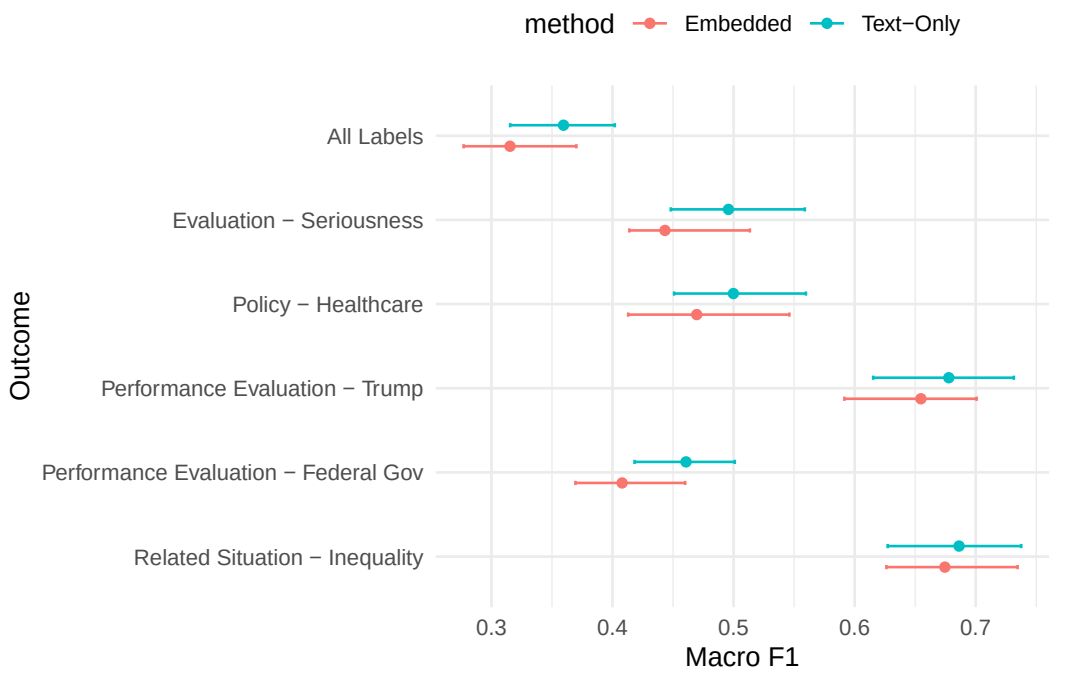

Figure 7. Comparing Performance of Machine Classifiers (Macro F1). The figure shows tweets that coders labeled with access to media content do not train better classifiers. Note: (1) Please refer to the Research Design section for the definitions of the outcomes/ tasks; (2) The Credible Intervals are obtained through bootstrapping the data 100 times for the treated and control groups respectively.

were to use a more sophisticated classifier that can incorporate both the text and the embedded data. Demonstrating empirically that this is the case, though, is beyond the scope of this study.

3. Conclusion

In this paper, we investigate how to improve human annotation of social media posts. We focus on the effect of giving coders access to posts' media content on the quality of labeling. We conduct an experiment on a tweet labeling task run in the fall of 2020. We design a web application for tweet labeling and randomly assign coders to a treatment group that has access to media content in tweets, and a control group where only the raw text of tweets are rendered. Our results show that giving human coders access to the media content of tweets appears to only marginally improve the quality of human annotation. We also note that access to media content led to substantial reduction of missing labels, and a slight improvement in the coders' job satisfaction, although these benefits come at the cost of time: coders take approximately two additional seconds to label a tweet in the

embedded media condition.

However, on the more important measure – the differences in classifier performance – we find that access to media content has an overall negative effect on the performance of machine learning classifier trained on such data.

We note that this result is potentially sensitive to the domain of interest. We posit that, where media and text information is aligned, media-enriched labeling can improve the quality of downstream tasks by allowing coders to more accurately capture the underlying message of the text. However, where media and text information are not aligned (for example, where sarcasm and irony are more prominent norms of communication), we posit that media-enriched labeling might hurt the quality of the overall task. We leave a more thorough investigation of this theory to future work.

Development of supervised machine learning classifiers relies on high quality training data sets. No matter what underlying technology is used in the classifier, the quality of the training data remains paramount. Determining the most efficient way to create such training data sets remains an important part of improving over all machine learning results. As researchers experiment with many ways to label data, our analysis offers a nuanced perspective on how to augment the human component of the supervised learning pipeline.

Funding Statement This work was supported by a grant from the Russell Sage Foundation. The Center for Social Media and Politics at New York University is generously supported by funding from the National Science Foundation, the John S. and James L. Knight Foundation, the Charles Koch Foundation, the Bill and Melinda Gates Foundation, the Hewlett Foundation, Craig Newmark Philanthropies, the Siegel Family Endowment, and NYU's Office of the Provost. This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

Competing Interests None.

References

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

King, Gary, Patrick Lam, and Margaret E Roberts. 2017. Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science* 61 (4): 971–988.

Liu, Yinhan, 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.* https://doi.org/10.48550/ARXIV.1907.11692. https://arxiv.org/abs/1907.11692.

Miller, Blake, Fridolin Linder, and Walter R Mebane. 2020. Active learning approaches for labeling text: review and assessment of the performance of active learning approaches. *Political Analysis* 28 (4): 532–551.

Wu, Patrick Y, Richard Bonneau, Joshua A Tucker, and Jonathan Nagler. 2022. Dictionary-assisted supervised contrastive learning. *arXiv preprint arXiv:2210.15172.*

Appendix 1. Full Description of the Annotation Task

We design a codebook, shown in Table A1, 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 A1. 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
	Election
Government performance	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 2. 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 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 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 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.

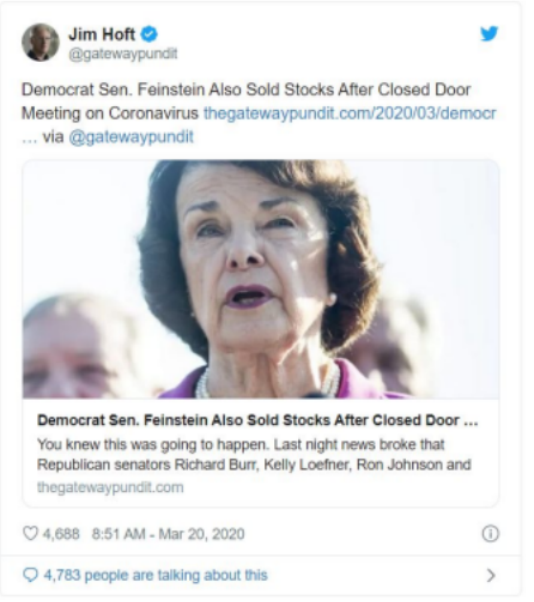

Figure 11. A tweet that shares questionable information

Figure 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 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 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 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 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.

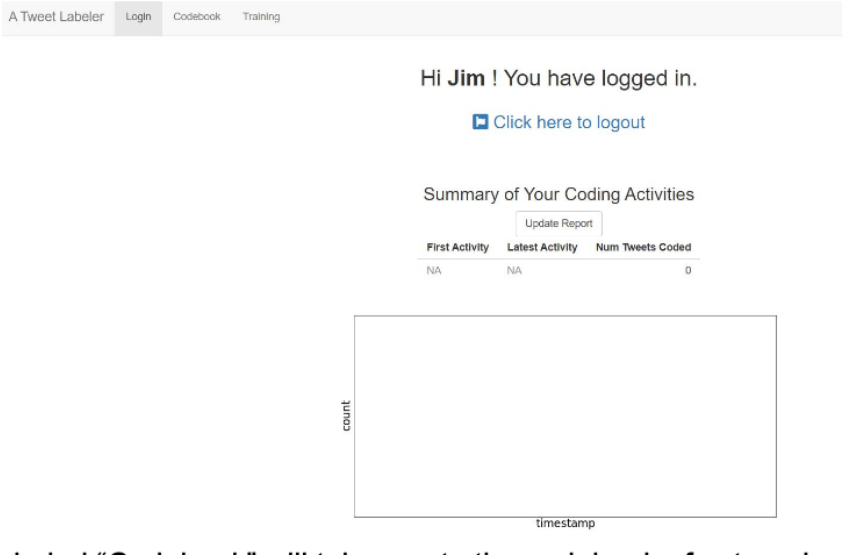

Figure 17. Login page of the labelling app

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.

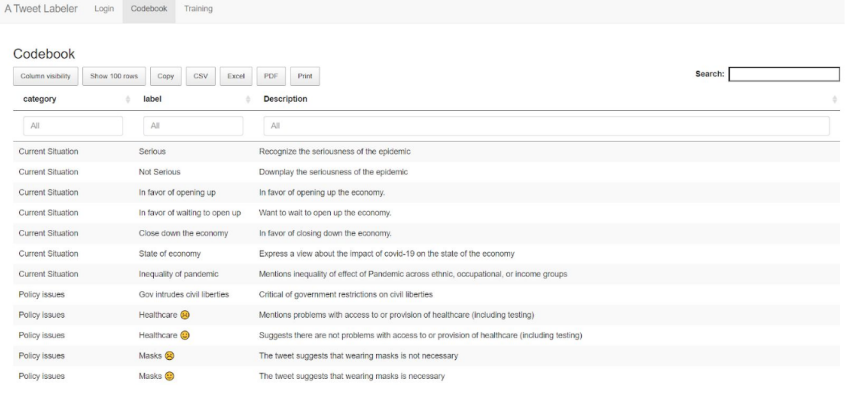

Figure 18. Codebook page in the labelling app

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 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.

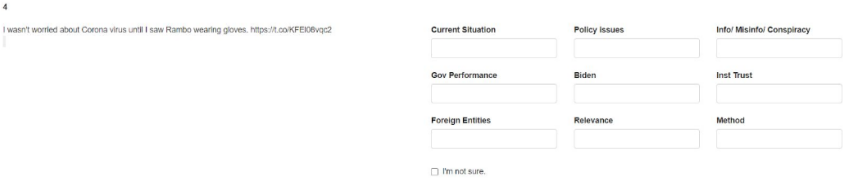

Figure 20. Example case when the embedded tweet fails to load

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.

Appendix 3. List of All Labels

ID	category	label	Description
1	Current Situation	Serious	Recognize the seriousness of the epidemic
2	Current Situation	Not Serious	Downplay the seriousness of the epidemic
3	Current Situation	In favor of opening up	In favor of opening up the economy.
4	Current Situation	In favor of waiting to open up	Want to wait to open up the economy.
5	Current Situation	Close down the economy	In favor of closing down the economy.
6	Current Situation	State of economy	Express a view about the impact of covid-19 on the state of the economy
7	Current Situation	Inequality of pandemic	Mentions inequality of effect of Pandemic across ethnic, occupational, or income groups
8	Policy issues	Gov intrudes civil liberties	Critical of government restrictions on civil liberties
9	Policy issues	Healthcare Disapprove	Mentions problems with access to or provision of healthcare (including testing and PPE)
10	Policy issues	Healthcare Approve	Suggests there are not problems with access to or provision of healthcare (including testing and PPE)
11	Policy issues	Masks Disapprove	Suggests that wearing masks is not necessary
12	Policy issues	Masks Approve	Suggests that wearing masks is necessary
13	Policy issues	Social Distancing Disapprove	Suggests 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.

List of labels continued from the previous page

ID	category	label	Description
14	Policy issues	Social Distancing Approve	Suggests approval 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.
15	Policy issues	School Closure Disapprove	Suggests disapproval of closing schools (or, opening them if closed)
16	Policy issues	School Closure Approve	Suggests approval of closing schools (or, keeping them closed longer)
17	Policy issues	Church Closure Disapprove	Suggests disapproval of closing churches (or, opening them if closed)
18	Policy issues	Church Closure Approve	Suggests approval of closing churches (or, keeping them closed longer)
19	Policy issues	Public Space Closure Disapprove	Suggests disapproval of closing public spaces (or, opening them if closed). Public spaces inclde beaches, playgrounds, and parks.
20	Policy issues	Public Space Closure Approve	Suggests approval of closing public spaces (or, keeping them closed longer). Public spaces inclde beaches, playgrounds, and parks.
21	Policy issues	Economic Relief Disapprove	Suggests a negative view about how the government is handling economic relief (this could be stimulus checks, rent freeze, anything)
22	Policy issues	Economic Relief Approve	Suggests a positive view about how the government is handling economic relief (this could be stimulus checks, rent freeze, anything)
23	Policy issues	Election	Suggests anything about elections (delays, vote by mail, other)

List of labels continued from the previous page

ID	category	label	Description
24	Info/ Misinfo/ Conspiracy	Factual information	Contains factual information relevant to the pandemic (but NOT anecdotes – "I got covid")
25	Info/ Misinfo/ Conspiracy	Information - unsure	Contains information relevant to the pandemic which you are not sure if it is true or false
26	Info/ Misinfo/ Conspiracy	Misinformation	Contains scientific or policy-related misinformation
27	Info/ Misinfo/ Conspiracy	Promotes a conspiracy theory	A conspiracy theory means an explanation of the current or past situation of the covid-19 outbreak as coordination among multiple people or a government with intention to deceive and/or remain anonymous.
28	Gov Performance	Federal	Mentions the federal government, or any actor within the federal govt, without a sentiment
29	Gov Performance	Federal Disapprove	Expresses negative sentiment about how the federal government, or any actor(s) within the federal govt is handling the covid-19 crisis.
30	Gov Performance	Federal Approve	Expresses positive sentiment about how the federal government, or any actor(s) within the federal govt is handling the covid-19 crisis.
31	Gov Performance	Trump	Mentions Trump without a sentiment
32	Gov Performance	Trump Disapprove	Expresses negative sentiment about Trump's handling of the Covid-19 crisis
33	Gov Performance	Trump Approve	Expresses positive sentiment about Trump's handling of the Covid-19 crisis
34	Gov Performance	Governor	Mentions governor(s) of any state(s) without a sentiment
35	Gov Performance	Governor Disapprove	Expresses negative sentiment about governor(s) of any state(s)'s handling of the covid-19 crisis

List of labels continued from the previous page

ID	category	label	Description
36	Gov Performance	Governor Approve	Expresses positive sentiment about governor(s) of any state(s)'s handling of the covid-19 crisis
37	Gov Performance	State or local policy	Mentions policies of any state(s) (or locality) without a sentiment
38	Gov Performance	State or local policy Disapprove	Expresses negative sentiment about the covid-19 related policies of any state(s) or locality
39	Gov Performance	State or local policy Approve	Expresses positive sentiment about the covid-19 related policies of any state(s) or locality
40	Biden	Biden	Mentions of Biden without any sentiment
41	Biden	Biden Disapprove	Expresses negative sentiment about Biden
42	Biden	Biden Approve	Expresses positive sentiment about Biden
43	Inst Trust	CDC	Mentions CDC without a sentiment
44	Inst Trust	CDC Disapprove	Expresses negative sentiment about CDC
45	Inst Trust	CDC Approve	Expresses positive sentiment about CDC
46	Inst Trust	Experts	Mentions about any high profile expert or scientist (Fauci, etc.)
47	Inst Trust	Experts Disapprove	Expresses negative sentiment about any high-profile expert or scientist (Fauci, etc.)
48	Inst Trust	Experts Approve	Expresses positive sentiment about any high-profile expert or scientist (Fauci, etc.)
49	Inst Trust	WHO	Mentions the World Health Organization (WHO) without a sentiment
50	Inst Trust	WHO Disapprove	Expresses negative sentiment about World Health Organization (WHO)
51	Inst Trust	WHO Approve	Expresses positive sentiment about World Health Organization (WHO)
52	Inst Trust	Media	Mentions Media
53	Inst Trust	Media Disapprove	Expresses negative sentiment about any media outlet or figure
54	Inst Trust	Media Approve	Expresses positive sentiment about any media outlet or figure
55	Foreign Entities	China	Mentions China

List of labels continued from the previous page

ID	category	label	Description
56	Foreign Entities	China Disapprove	Expresses negative sentiment about China
57	Foreign Entities	China Approve	Expresses positive sentiment about China
58	Foreign Entities	Europe	Mentions Europe
59	Foreign Entities	Europe Disapprove	Expresses negative sentiment about Europe
60	Foreign Entities	Europe Approve	Expresses positive sentiment about Europe
61	Foreign Entities	Russia	Mentions Russia
62	Foreign Entities	Russia Disapprove	Expresses negative sentiment about Russia
63	Foreign Entities	Russia Approve	Expresses positive sentiment about Russia
64	Foreign Entities	Other	Mentions other foreign entities
65	Foreign Entities	Other Disapprove	Expresses negative sentiment about other foreign entities
66	Foreign Entities	Other Approve	Expresses positive sentiment about other foreign entities
67	Bias or Hate Speech	Asian-Americans Disapprove	Expresses a negative sentiment towards Asian-Americans
68	Bias or Hate Speech	Asian-Americans Approve	Rejects blaming Asian-Americans in the context of Covid-19
69	Bias or Hate Speech	Immigrants Disapprove	Expresses a negative sentiment towards Immigrants
70	Bias or Hate Speech	Immigrants Approve	Rejects blaming Immigrants in the context of Covid-19
71	Bias or Hate Speech	Other groups Disapprove	Expresses a negative sentiment towards other groups (other than immigrants or Asian-Americans)
72	Bias or Hate Speech	Other groups Approve	Rejects blaming other groups in the context of Covid-19 (other than immigrants or Asian-Americans)
73	Relevance	Covid-19-OTHER	The tweet is relevant to Covid-19 in the United States, but does not fit in any other category we provided
74	Relevance	Irrelevant - not COVID-19	The tweet is NOT about COVID-19.
75	Relevance	Irrelevant - not the US	The tweet is about Covid-19. But it is not relevant to COVID-19 in the United States, OR is not by an individual in the United States.

List of labels continued from the previous page

ID	category	label	Description
76	Relevance	Not enough information	The tweet does not provide enough information for coding.
77	Method	Image	Relied on an Image with the Tweet
78	Method	Video	Relied on a Video with the tweet
79	Method	Followed a Link	Relied on content in a link in the tweet