Political Science Research and Methods (2026), **14**, 1019–1031
doi:10.1017/psrm.2025.10010

RESEARCH NOTE

Labeling social media posts: does showing coders multimodal content produce better human annotation, and a better machine classifier?

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

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

(Received 30 January 2024; revised 18 November 2024; accepted 17 January 2025; first published online 2 July 2025)

Abstract
The increasing multimodality (e.g., images, videos, links) of social media data presents opportunities and challenges. But text-as-data methods continue to dominate as modes of classification, as multimodal social media data are costly to collect and label. Researchers who face a budget constraint may need to make informed decisions regarding whether to collect and label only the textual content of social media data or their full multimodal content. In this article, we develop five measures and an experimental framework to assist with these decisions. We propose five performance metrics to measure the costs and benefits of multimodal labeling: average time per post, average time per valid response, valid response rate, intercoder agreement, and classifier's predictive power. To estimate these measures, we introduce an experimental framework to evaluate coders' performance under *text-only* and *multimodal* labeling conditions. We illustrate the method with a tweet labeling experiment.

Keywords: mass media and political communication; measurement; text and content analysis

1. Introduction

Computational social scientists have extensively used social media posts to measure the political opinions of politicians, media outlets, and the public (e.g., Barberá *et al.*, 2019). In these efforts, researchers typically use a machine-assisted approach to content analysis. Most existing political methodology studies on machine-assisted content analysis have focused on text. Political scientists have developed sophisticated text-as-data methods, including efficient algorithms and best practices, for the retrieval (King *et al.*, 2017), manual labeling (Barberá *et al.*, 2016; Benoit *et al.*, 2021), and machine-learning classification (Grimmer and Stewart, 2013; Grimmer *et al.*, 2021) of political texts. Lately, political scientists have incorporated large language models for machine-assisted labeling of social media posts (Gilardi *et al.*, 2023). However, by ignoring the non-textual features of these contents, scholars risk ignoring or even misunderstanding the meaning of the text.

We start from the observation that multimodal contents of social media data can impact the interpretation and, subsequently, labeling of social media posts. Multimodal content includes images, videos, and links that authors include as parts of social media posts. It also includes metadata of the authors of social media posts (e.g., profile pictures, self-reported locations), as well as the design

© The Author(s), 2025. Published by Cambridge University Press on behalf of EPS Academic Ltd. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (http://creativecommons.org/licenses/by-nc-nd/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided that no alterations are made and the original article is properly cited. The written permission of Cambridge University Press must be obtained prior to any commercial use and/or adaptation of the article.

https://doi.org/10.1017/psrm.2025.10010 Published online by Cambridge University Press

of the platform's user interface. All else equal, multimodal content can improve the ability of human coders to accurately label social media data according to the quantities of substantive or theoretical interest.[1]

Existing studies on multimodal social media data focus on *how* to conduct these studies *if* researchers have the need and the data. These methodological inquiries focus on effective feature engineering and machine learning modeling (Peng *et al.*, 2024; Peng *et al.*, 2023; Wu and Mebane, 2022). Despite these recent methodological advancements, incorporating multimodality incurs additional costs, which likely explains the slow pace of its adoption in applied political science research. Researchers who analyze multimodal social media data with a constrained budget may need to make an informed decision regarding whether to incorporate the multimodality or use only their extracted text.

In an ideal scenario, it is always advisable to incorporate the multimodality of social media data into the process of data annotation. We assume the objective of content analysis is to create valid measures to accurately reflect the information conveyed by the social media posts, or the intention of the authors of the posts, depending on the theory of interest. Incorporating multimodal content, which is an integral attribute of the online information, through the labeling and classifying process can generally improve the quality of the measures as both coders and models access information that is close to what the online audience will access.

However, the ideal scenario is not always feasible because multimodal social media data are *costly* to collect, label, and model. Budget constraints associated with data access, human labor, and computational resources can force researchers to choose between using only the textual contents of social media posts or the full multimodal contents at both the labeling and classifying stage. First, multimodal content can be more costly to collect and store. For example, researchers may need to purchase Application Programming Interface (APIs) with higher capacity from social media platforms to collect non-text content, develop and deploy web scrapers that can handle multimedia data, and/or use more complicated database structures to read and write multimedia data. Second, multimodal content can be more costly to manually label. Cleaning and labeling these more complex multimodal data typically requires more hours of human labor (e.g., crowd-sourced workers and research assistants) and higher unit prices. Finally, multimodal content can be more costly to label by machines. Larger deep neural network models need to be trained and applied to multimodal data. Even with the recent breakthrough of large language models (see Zhang and Pan (2019) and Wu and Mebane 2022 for recent examples), the costs (e.g., computing time, data input and output pipeline, and pay for commercial APIs) required to handle multimodal content are still likely to be significantly higher than those for text-only content. Practically speaking, the vast majority of applied political science research in this area relies on text-based classifiers, underscoring the baseline importance of our contribution.

Facing a budget constraint, how can researchers make an informed decision regarding whether to use only the textual content of social media posts ("text-only labeling" hereafter) or incorporate their multimodal content ("multimodal labeling" hereafter) when they are limited by budget or technical constraints to using a text-based classifier in the downstream task of classifying their entire corpus? We argue that the answer lies in an empirical evaluation of the extent to which multimodal labeling, compared to text-only labeling, changes (1) the human coders' interpretation and labeling of the content, and (2) text-based machine learning classifiers' downstream performance. In this paper, we introduce a study design to facilitate this empirical evaluation. The design contains five performance metrics and an experimental framework. We illustrate the design with a crowd-sourced tweet-labeling project to measure Americans' public opinion on COVID-19 in 2020.

We start from the assumption that multimodal features are informative of meaning, and the assertion that any data annotation exercise is fundamentally an effort to accurately encode the underlying

[1]Recently, an emerging interdisciplinary literature has applied computational methods to incorporate multimodal content for the analysis of social media content. These include, for example, detecting protest events with images (Zhang and Pan, 2019) and identifying effective fact-checking strategies with videos (Hameleers *et al.*, 2020; Lu and Shen, 2023).

https://doi.org/10.1017/psrm.2025.10010 Published online by Cambridge University Press

Table 1. Measures for the cost and benefit of multimodal labeling compared to text-only labeling

Performance metrics	Measure	Interpretation	Stage
Time taken per post (T)	$\Delta T = T(\text{multi}) - T(\text{text})$	Larger $\rightarrow$ costs $\uparrow$	First
Time taken per valid response (T_v)	$\Delta T_v = T_v(\text{multi}) - T_v(\text{text})$	Larger $\rightarrow$ costs $\uparrow$	First
Valid response rate (R)	$\Delta R = R(\text{multi}) - R(\text{text})$	Larger $\rightarrow$ benefits $\uparrow$	First
Intercoder agreement (I)	$\Delta I = I(\text{multi}, \text{multi}) - I(\text{multi}, \text{text})$	Larger $\rightarrow$ benefits $\uparrow$	First
Classifier predictive power (P)	$\Delta P = P(\text{multi}) - P(\text{text})$	Larger $\rightarrow$ benefits $\uparrow$	Second

Note: "multi" is short for multimodal labeling and "text" is short for text-only labeling.

meaning, attitude, intent, or argument of the individuals who posted the content or the interpretation thereof by the audience for these posts. All else equal, giving human annotators access to these features should improve their ability to infer the meaning of a piece of content, thereby applying a more accurate label that reflects the underlying concept(s) of theoretical interest. However, these labels might not necessarily improve the performance of a downstream classifier if the classifier does not also have access to the multimodal features, as is the case across the vast majority of applications in computational social science research. While we disaggregate these costs and benefits in more detail in our subsequent analyses, it is worth underscoring that our core investigation is into whether the theorized improvements to human annotation afforded by access to multimodal content translate into better downstream classification when the classifier only has access to the text. Throughout, the overarching goal is to accurately annotate data. All else equal, we expect the human annotations to improve with access to multimodal content. However, we suspect that the text-based machine learning classifiers might show no improvement or even a decline in performance when trained with these mutlimodal-labeled data.

2. Measuring the costs and benefits of multimodal labeling

We define the standard data annotation exercise as a two-stage procedure. In the first stage, human coders annotate a subset of the data. In the second stage, these data are used as training data for fine-tuning a downstream classifier which can then be used to dramatically scale up the amount of annotated data.[2] We propose five performance metrics to measure the costs and benefits of multimodal labeling, divided into four metrics that speak to human annotation performance in the first stage, and a final metric that speaks to classifier performance in the second stage. The metrics are summarized in Table 1 and detailed in the remainder of this section.

Our first performance metric assesses whether multimodal labeling requires more human labor than text-only labeling. We operationalize this concept as the average time taken (T) to label a post. For the same labeling task, the difference between the time taken to label it with multimodal information $T(\text{multi})$ and that taken for text-only information $T(\text{text})$ is the *cost* of multimodal labeling (ΔT). We are agnostic about the direction of this metric. On the one hand, access to multimodal content might require more time for the human annotator to parse these non-textual features before deciding on a label. On the other hand, access to multimodal content might make the interpretation task easier, reducing the time taken to label.

Our second and third performance metrics assess how multimodal labeling influences coders' capacity to understand the meaning of the content: the average time taken to give a valid response (i.e., not indicating N/A or "not enough information") and the valid response rate. Expressions on social media can be informal, ambiguous, and contextualized. In some cases, coders can find the textual content insufficient to inform their coding decisions. Multimodal labeling may provide coders with contextual information to make a decision. For example, links, images, and videos can provide background information about the political events or persons of interest a post is referring to. When

[2]We are not interested in the cases where either the total data to be annotated are sufficiently small or the researcher is sufficiently unconstrained such that human annotation is feasible for the full dataset. In this setting, it is straightforward to recommend using multimodal information to improve the accuracy of the human labels.

https://doi.org/10.1017/psrm.2025.10010 Published online by Cambridge University Press

the text states "he is corrupt," the associated image may reveal which politician "he" refers to. The authors' profile can disambiguate the political entity of interest. For a post commenting on "our governor," the author's geographic location can clarify to which governor it refers. We operationalize the average time taken for a valid response as the total time taken for the labeling task divided by the total number of valid responses. We operationalize the valid response rate as 1 minus the proportion of posts assigned the "don't know" label. For the same labeling task, the difference between the proportion of posts for which coders return valid labels under multimodal labeling [R(multi)] and that under text-only labeling [R(text)] is a first benefit of multimodal labeling.

Our fourth performance metric assesses how often coders disagree on the appropriate label in the multimodal condition compared to those in the text-only labeling condition. We operationalize this metric as the difference between two measures of intercoder agreement among coders labeling the same posts under different labeling conditions: the intercoder agreement between two coders who both access multimodal information (I(multi, multi)), and the intercoder agreement between a coder accessing multimodal information and a coder accessing text-only information (I(multi, text)). The difference between the two, $\Delta I = I(\text{multi, multi}) - I(\text{multi, text})$, provides measures for a second benefit of multimodal labeling.

This performance metric (ΔI) should be interpreted with two caveats. First, we are interested in the intercoder agreement in a relative, not absolute, sense. Conventionally, intercoder agreement measures the quality of coding tasks, with higher values generally meaning better coding quality. Our focus is not on how high they are. Rather, we are primarily interested in the *differences* between the two intercoder agreement measures. These differences measure how often taking away multimodal content (while leaving textual content) changes coders' decisions. The first part of the equation, [I(multi, multi)] serves as a benchmark, which can account for the complexity of the labeling tasks. In this way, ΔI is comparable across different parts of the same labeling task and even different labeling tasks. Second, the underlying assumption of our operationalization is that the labels generated from multimodal labeling are our "gold standard." Considering them the gold standard does not mean coders labeling under multimodal conditions always return the correct answer. Instead, we consider them the gold standard because we assume they return the best achievable labeling performance given the same content, personnel, and equipment. This is consistent with our overarching assumption that, all else equal, multimodal content provides more information about the meaning of a post, and that the quality of the labeling is increasing in the amount of information. In the analyses that follow, we treat the labels from the multimodal task as the gold standard since they best-approximate our theoretical quantity of interest: the meaning of the post itself.

Our fifth performance metric turns to the second stage of the data labeling exercise and assesses the predictive capacity of machine learning classifiers trained on the data labeled under multimodal and text-only conditions, respectively. When the number of social media posts to study exceeds the human capacity to label them, researchers typically manually label a portion of the posts and then train a machine learning classifier to label the remainder. While the first four performance metrics assess the impact of multimodal content on the manual labeling part of the content analysis, this final metric assesses its impact on machine classification. The difference between the predictive power of classifiers trained with multimodal-labeled data [P(multi)] and those trained with text-only data [P(text)], $\Delta P = P(\text{multi}) - P(\text{text})$, indicates the extent to which multimodal content helps train better machine classifiers.

As discussed above, we are interested in whether the inclusion of multimodal content at the labeling stage can improve downstream classification performance. Importantly, our evaluation of this research question holds constant the *type* of classifier we use. In both the multimodal and text-only analyses, we use the same Bidirectional Encoder Representations from Transformers (BERT)-based algorithm to predict the labels. These classifiers differ only in the training data that they are fine-tuned on, and here only in the sense that the same training data have labels applied in different information environments. In other words, the same algorithm is given the same text features for the same social

https://doi.org/10.1017/psrm.2025.10010 Published online by Cambridge University Press

media posts, but we train one on the posts which were labeled in the multimodal condition, while the other is trained on the posts which were labeled in the text-only condition.

We are agnostic as to whether training data labeled in the multimodal condition will produce better or worse downstream classifier performance when the downstream classifier only has access to the text features of the training data. On the one hand, we might expect that multimodal labels are more *accurate* because the human annotators have access to more information about the meaning of a given post. For example, when a politician is systematically referred to by a nickname unfamiliar to coders, coders can use accompanying multimodal contexts (e.g., photos of the politician and links to news reports about the politician) to associate the posts with the politician correctly. The machine classifier, on the other hand, despite its lack of access to the multimodal context, might be able to learn from the textual content to build an association between the nickname and the politician to make correct out-of-sample predictions.

Conversely, training data labeled in the multimodal condition relies on non-textual features that the downstream text-based classifier can't access. As such, training data from the multimodal labeling condition might instead confuse the classifier. For example, a multimodal post might contain only the text "see this video!", along with an embedded video. Although the label might correctly reflect the pandemic-concerned nature of the video, the text itself is uninformative of this label. The classifier, which can only access text, can identify "video" as a predictor of the posts' relevance to COVID-19 and mislabel other posts referring to videos that are irrelevant to the topic.[3]

To measure the predictive power of the downstream classifier, we use conventional performance metrics for machine learning modeling, including *accuracy*, *precision*, *recall*, *F1* of out-of-sample predictions. Similar to the third performance metric, we define the "gold standard" for out-of-sample predictions as the data manually labeled under the multimodal condition. We assert that these labels are superior reflections of the underlying quantity of interest: the meaning intended by the post's author and, consequently, how it will be interpreted by its audience.

In sum, researchers can consider the above five performance metrics to decide whether to incorporate multimodal content in machine-assisted content analysis with social media posts. Depending on the research questions and the specifics of budget constraints, researchers may weigh the five performance metrics differently in the decision-making process. For example, researchers with scarce human labor for data labeling may be more sensitive to the additional time required for multimodal labeling than others.

3. An experimental framework for cost–benefit analysis

We develop an experimental framework to estimate the five performance metrics. We consider this framework applicable to the pilot stage of any labeling job, when researchers evaluate the costs and benefits of incorporating multimodal information. The experiment can help researchers gather information using a sample of the data before conducting analyses at scale. The crux of this experimental framework is constructing and evaluating a sample of quadruple-coded social media posts. The process takes four steps.

First, researchers randomly sample a small proportion of the social media posts requiring annotation. Second, central to this design, each post in the sample is randomly assigned to *four* coders. Two

[3]While not the focus of our study, researchers can develop a multimodal classifier that can use both the text-based and multimodal features as inputs and compare it with the text-based classifiers we analyze here. This fully multimodal annotation pipeline is likely optimal, as the machine accesses the richest set of data, approximating human social media consumers. However, building such models can be costly and even infeasible—the very reason why budget-constrained researchers may need our proposed assessment. Furthermore, advances in these types of algorithms are very recent and still largely untested, particularly those that exploit the enormous strides made in large language models (although see Zhang and Pan (2019) and Wu and Mebane 2022). Hence, we do not include multimodal classifiers in our evaluation metrics in this study.

https://doi.org/10.1017/psrm.2025.10010 Published online by Cambridge University Press

of these coders, considered the *treated* group, are instructed to label the posts with multimodal information. The other two coders, considered the *control* group, are instructed to label the post based only on their textual content.

After completing the labeling task, the researcher should then obtain the performance metrics defined in Section 2. First, researchers can use the quadruple-coded posts to calculate the first through the fourth performance metrics regarding the costs and benefits associated with the manual labeling stage. Then, researchers can fit two machine learning classifiers: one with multimodal labeled data and one with text-only labeled data. *Both* models use only the textual features of the same posts as predictors but differ in whether the labels were assigned by human coders with access to the multimodal content or not. The two models can then be evaluated to calculate the fifth performance metric regarding the predictive power of the machine learning classifiers.

4. Illustration: a tweet labeling experiment

We illustrate our performance metrics and experimental framework with a tweet labeling task conducted in the summer and fall of 2020[4]. We recruited 12 coders to label tweets for a research project on public sentiment about the COVID-19 pandemic. The coders were instructed to code the tweets for whether they contain discussions on one or more of the following topics:

- **Evaluation of COVID-19 seriousness**: Whether the tweet take COVID-19 in the United States *seriously* or *not seriously*.
- **Concerns about the economic consequences of COVID-19**: whether the tweet expresses concerns about the economy (in favor of opening up or waiting to open up the economy); and whether a tweet expresses concerns about inequality caused by COVID-19.
- **Attitudes toward COVID-19 policies**: whether the tweet expresses attitudes (approval, neutral, disapproval) toward three policy issues: healthcare policies, mask-wearing requirements, and economic relief.
- **Political support related to the handling of COVID-19**: whether the tweet expresses attitudes (approval, neutral, disapproval) toward the federal government, the president, and governors with reference to their handling of COVID-19, respectively.

4.1. Experimental setup

We split the labeling tasks into five assignments of equal size. Within each assignment, half of the coders had access to only the textual contents of the tweets (the "text-only" group), while the other half could access both the tweets' text content as well as the multimodal content (the "multimodal" group). Coders alternated between the two groups across weekly assignments. For example, if a coder was in the text-only group in the first assignment, they were in the multimodal group in the second assignment, and then switched back to the text-only group in the third assignment.[5] We allowed coders to exploit the multimodal information as they saw fit to understand the meaning of the post, including clicking on any hyperlinks. In practice, coders chose to do this in only a small fraction of cases. Weekly debriefs indicate that they chose to do so primarily to better label the post or, in very few cases, because they were curious about the information. See Appendix 1 for the details.

Our team of coders was tasked with labeling a total of 12,026 tweets. The coders were randomly assigned to label these tweets either: 1) based on their textual information only (the "text-only"

[4]Our experiment was not preregistered. As such, we consider all the subsequent results exploratory.

[5]We change the treatment assignments for our human coders on a weekly basis to improve precision and avoid small-sample bias. We don't believe there should be confounding via SUTVA violations as coders shift from the text-only to the multimodal condition, or in reverse, week-to-week.

https://doi.org/10.1017/psrm.2025.10010 Published online by Cambridge University Press

group); or 2) based on full multimodal information and content (the "multimodal" group) . By the end of the experiment, we constructed two datasets out of the labeled tweets. The first dataset includes 2,351 *quadruple-coded* tweets, each labeled by two coders in the multimodal condition and two coders in the text-only condition. We use these quadruple-coded tweets to estimate the first four performance metrics regarding human labeling, including time (T and T_v), valid response rate (R), and intercoder agreement (I). The second dataset includes 7,914 *double-coded* tweets, each labeled by one coder in the multimodal condition and one coder in the text-only condition. We combine the quadruple-coded and double-coded tweets from the multimodal condition to estimate the performance metrics of the machine learning classifiers.[6,7] We evaluate performance on *only* labels from the multimodal condition because we believe these more accurately reflect the underlying quantity of interest. Given this, one might expect that the classifier trained on the multimodal-labeled data should perform better, since its training data are more similar to its test data. As we demonstrate below, this is not the case.

4.2. Experimental infrastructure

To implement this experiment in a way that maximizes compliance to treatment, we designed a web application using *R Shiny*. When coders used this application to label the tweets, they saw a two-column layout: the tweets' contents were in the left column, while the labels to apply were in the right column. The multimodal and text-only groups saw tweets' contents displayed differently in the left column. Figure 1 shows the difference with an example. Panels (a) and (b) show how the web application appeared for coders in the text-only and multimodal groups, respectively. The text-only group could read only the plain text extracted from the tweets. In contrast, the multimodal group saw the extracted text *and* an embedded window showing how the tweet appeared on the Twitter timeline, including the authors' profile information, engagement metrics, links, and media (if any).[8]

4.3. Findings

We estimate the performance metrics using the data collected from this experiment. Figure 2 provides an overview of the effect of multimodal labeling in comparison to text-only labeling. We find that multimodal labeling increased the time spent labeling each post by 19% (ΔT) compared to text-only labeling and increased the time spent per valid response (ΔT_v) by 14%. Despite these costs in terms of time, multimodal labeling increased the valid response rate (ΔR) by 4%. In addition, the intercoder agreement (ΔI) between two coders who both labeled with access to multimodal content is almost 10% higher than that between a coder with multimodal access and another with text-only access. This suggests that multimodal labeling significantly changed the coders' interpretation and labeling of the posts.

In sum then, there is clear evidence that multimodal labeling significantly affects the first four performance metrics that speak to the first stage of the classification process. While the human coders took significantly more time to label the same data, they had fewer posts that they weren't able to label, and achieved a higher intercoder agreement. This first set of findings confirms our assertion

[6]It would be ideal to quadruple-code all tweets. This combination is a workaround to increase the number of unique tweets labeled.

[7]Using the labeled data, we estimate a text-only classifier and a multimodal classifier. *Both* models use textual features as predictors. For outcomes, the text-only classifier uses labels generated under the text-only labeling condition, while the multimodal classifier uses labels generated under the multimodal labeling condition. For the rationale of this design, see Section 2.

[8]The tweets were embedded using Twitter's oEmbed API.

https://doi.org/10.1017/psrm.2025.10010 Published online by Cambridge University Press

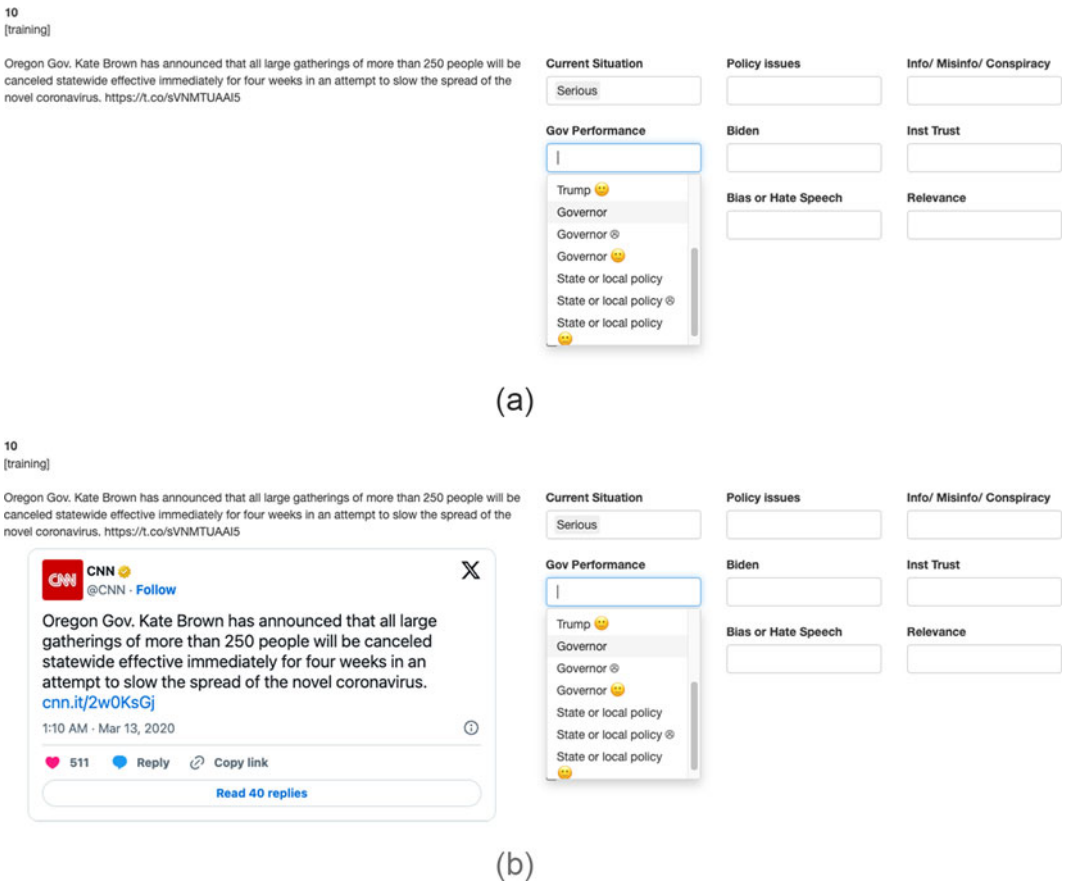


Figure 1. Interface of our tweet labeling infrastructure.

that the multimodal condition yields "better" labels in the sense that they more accurately reflect the underlying quantity of theoretical interest: the meaning of, and information contained, in the tweet. However, the additional time required to code also confirms our assertion that multimodal labeling is more costly.

But does this translate into better performance for the a downstream text-based classifier? As mentioned above, we might expect this performance to go in one of two ways. On the one hand, because the text-based classifier does not have access to the non-text features of the multimodal data, its performance might be lower if it gets confused. Conversely, if the text is sufficiently rich and the classifier is sufficiently sophisticated, it might nevertheless learn how the text is associated with non-text features. Finally, all else equal, we might think that the multimodal trained classifier performs better simply because the data on which it is evaluated is from the human coders in the multimodal condition.

Our analysis indicates that the text-based classifier performs slightly worse when trained on data labeled in the multimodal condition. The predictive power (ΔP) of the machine learning classifier trained with multimodal labeling is, on average, 3% lower than that of the classifier trained with text-only labeling (measured by the average F1 score), although this difference is not statistically significant at the 95% level of confidence. The remainder of this section elaborates on how we estimate each of these performance metrics.

https://doi.org/10.1017/psrm.2025.10010 Published online by Cambridge University Press

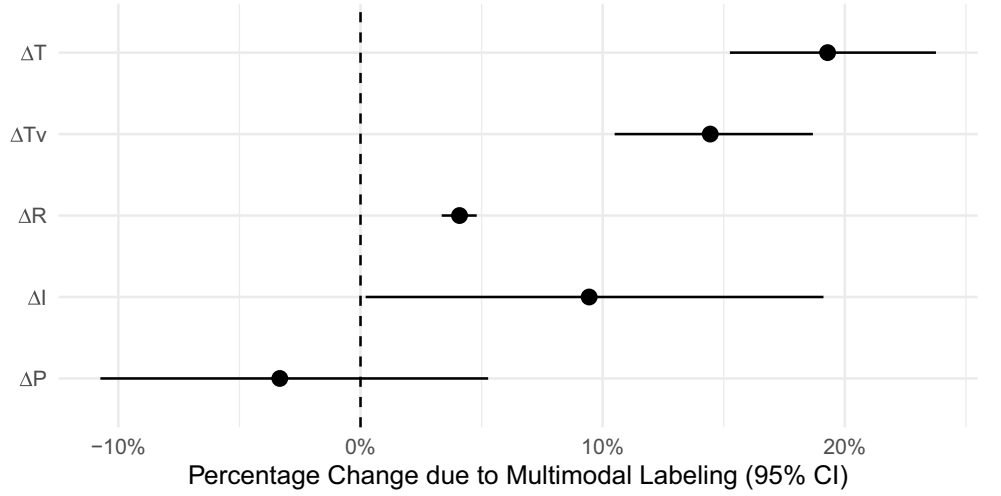


Figure 2. Multimodal labeling cost 19% more time per Twitter post and 14% more time per valid response, increased the valid response rate by 4%, and increased the intercoder agreement by 9%. It decreases classifier's predictive power by 3%. Horizontal bars indicate 95% intervals based on cluster robust bootstrapped standard errors for δT, ΔT_v, and ΔR. Inference for ΔI is based on bootstrapping two coders per tweet ID, while ΔP is based on 100 cross-validated calculations.

4.3.1. Multimodal labeling costs more time

Our web application logged the timestamp whenever a coder selected a label for a tweet. We use these timestamps to calculate the time it took coders to assign a label. Specifically, we calculate the time spent on a tweet as the timestamp associated with its labels minus the timestamp associated with the previous tweet labeled. Coders may split their work into different work sessions. We remove any durations that are greater to 5 minutes (which is higher than the 95th percentile of the distribution), as we consider these as signals that the two tweets are labeled in two different work sessions. We then estimate the differences between the time spent on a tweet for coders in the multimodal and text-only conditions. To fully account for the uncertainty associated with sampling, we approximate the estimate and confidence interval with bootstrapping. The bootstrapped average time for multimodal labeling is 26.6 seconds, while that for text-only labeling is 22.3 seconds. Multimodal labeling increased the time taken to label a tweet by an average of 19.3% (with a confidence interval from 15.3% to 23.8%). This suggests that multimodal labeling substantially increased the time required for the labeling task.

4.3.2. Multimodal labeling improved valid response rate

In our labeling instructions, we asked coders to apply a "not enough information" label when they were unable to obtain sufficient information to determine the meaning of a tweet. We measure the valid response rate as 1 minus the proportion of tweets that coders labeled as "not enough information." We then calculate the differences between the multimodal and text-only coders. We approximate the mean and confidence interval by bootstrapping. The mean valid response rate for the multimodal group is 94.4% and the for the text-only group is 90.7%. Multimodal labeling brings a statistically significant 4% increase in the valid response rate with a 95% confidence interval ranging from 3.4% to 4.8%. Substantively, this corresponds to 1,179 tweets assigned the "not enough information" label in the text-only condition, compared to 710 in the multimodal condition (see Appendix Table 4 for a detailed breakdown of tweet labels).

Below are a few examples of tweets that at least one coder flagged as not having enough information for labeling:

https://doi.org/10.1017/psrm.2025.10010 Published online by Cambridge University Press

- *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.*

All these examples highlight the importance of multimodal information. Two of them appear to refer to a linked web page or tweet. It is intuitive to expect that coders who can access these multimodal contexts through the embedded tweets in the coding app are in a better position to assign a valid label.

4.3.3. Multimodal labeling changed labels chosen

To quantify how much the multimodal content changed the substantive labels chosen by our coders, we create measures using our quadruple-coding setup (two labels based on multimodal content and two labels based on text-only content for each tweet). We compared two measures of intercoder agreement: (1) the intercoder agreement between two coders in the multimodal condition and (2) the intercoder agreement between a coder in the multimodal condition and a coder in the text-only condition. The more (1) exceeds (2), the more multimodal information changes the label chosen. As our labeling task contains multiple non-mutually exclusive categorical labels, we calculate Fleiss' kappa for each individual label respectively and then calculated their averages.[9] Notably, given our quadruple-coding setup with two coders in each group, there are four cross-condition intercoder agreements. We take averages as indicators. Let the two multimodal labels be *multi-1* and *multi-2* and the two text-only labels be *text-1* and *text-2*. Then (1) is $I(\text{multi-1}, \text{multi-2})$, while (2) is the mean of $I(\text{multi-1}, \text{text-1})$, $I(\text{multi-1}, \text{text-2})$, $I(\text{multi-2}, \text{text-1})$, $I(\text{multi-2}, \text{text-2})$.

The average intercoder agreement between two coders in the multimodal condition is 0.43, while that between a coder with access to multimodal content and a coder with text-only access is 0.39. Calculating the pair-wise relative differences of the bootstrapped sample, the former is 9.4% higher than that of latter, with a 95% confidence interval ranging from 0.2% to 19.1%. The difference shows that multimodal labeling had a significant impact on the substantive labels chosen by coders. All else equal, we expect the labels from the multimodal condition to be better reflections of the ground truth meaning of these posts.[10]

4.3.4. Multimodal labeling hurt the classifier's predictive power

Finally, we evaluate the impact of multimodal labeling on the predictive power of our machine learning classifier. We train two sets of machine learning classifiers: one using tweets labeled by coders who have multimodal access and the other using tweets labeled by coders with text-only access, respectively. We then evaluate out-of-sample performance using a held-out set of tweets labeled in the multimodal condition. Given that our labels are non-mutually exclusive categorical labels and our input is text data, we fit multi-label classifiers based on RoBERTa (Liu *et al.*, 2019). We use the average $F1$ scores as our measure of each classifiers' predictive power. We approximate the confidence interval through bootstrapping. Specifically, we create 100 bootstrapped datasets for the multimodal and text-only tweets, respectively. We then trained and evaluated the classifier for each of the bootstrapped datasets to obtain a sample of average $F1$ scores for both the treated and control groups. Importantly, the calculation of the $F1$ score relied on cross-validation with a 50–50 split in which we divided the bootstrapped sample at random (without replacement). The training half was either the multimodal or text-only annotated data. The test half was the set of tweets that did not appear in the training half but *were annotated by coders in the multimodal condition*. This decision reflects

[9] We treat missing labels as its own category when calculating Fleiss' kappa.

[10] Although we are interested in the difference in intercoder agreement metrics between the text-only and multimodal conditions, the Fleiss' kappa ranges from −1 to 1 where −1 indicates complete disagreement and 1 indicates complete agreement. A value of 0.43 has no direct substantive interpretation but is considered a poor-to-decent measure.

https://doi.org/10.1017/psrm.2025.10010 Published online by Cambridge University Press

our interest in evaluating the model with reference to the most accurately coded tweets, which we assume and demonstrate were those labeled by coders who had access to both text and non-text media.

The average $F1$ score trained with multimodal labeled tweets is 0.32, while that of text-only labeled tweets is 0.36. Calculating the pairwise differences with bootstrapped samples, the predictive power of the best classifier trained using labels generated under the multimodal labeling condition is 10% lower than that of the best classifier trained using labels generated under the text-only condition.[11] In the Appendix, we compare performance across a range of labels that appear frequently in the data, confirming that the penalty to the classifiers trained on the multimodal data persists even for labels where the classifier performance is substantially higher.

4.4. Discussion

In this tweet labeling task, the findings from the five performance metrics suggest that incorporating multimodal information in the labeling process brings about both costs and benefits. In terms of costs, human coders with access to multimodal information take significantly more time to annotate training data compared to those with only access to the raw text, measured as either time per labeled post or time per valid label. This means researchers need more human labor, which means increased costs.

In terms of benefits, multimodal labeling significantly increases the valid response rate. This means researchers have more valid data points for downstream analysis and classifier training. In addition, multimodal labeling significantly changes the substantive labels that coders would apply and improves intercoder reliability. These results suggest that the multimodal content carries information that the text systemically misses and hence significantly influences the coders' labeling decisions. Based on these results, we assert that multimodal content improves the accuracy of the human annotated data, which should be the first-order goal of any applied research.

However, the benefits of multimodal content in the human annotation stage did not translate into improvements in the text-based classifier. When trained on data labeled by coders with access to non-textual features, the same classifier yielded an *F1* score 10% lower than when it was trained on data labeled by coders with only access to the raw text. The finding suggests that the machine learning classifier cannot use textual heuristics to make better predictions. In fact, the multimodal training data may have produced misleading textual heuristics that confused the text-based classifier and led to worse performance.

Our results suggest caution when using additional information to manually label training data for downstream classifiers, if the downstream classifiers are unable to access all features used by the human coders. While the improvements to the number of valid labels, and the superior intercoder agreement, are attractive properties of multimodal labeling tasks, we find that these benefits do not carry forward to superior downstream classification performance. If anything, our analysis suggests that labels assigned with access to multimodal information can confuse a text-based classification algorithm. The first-order goal of text classification should be to accurately label content according to the theoretical quantities of interest. As such, the path forward should be to give human coders access to the multimodal content *and* adopt recent advances in machine learning algorithms that can exploit both the text and non-text features of the training data. However, our results suggest that if the immediate goal is to maximize performance of a text-based classifier, then labeling just based on text may be the best option.[12] In either case, our summary

[11] The difference may be considered statistically insignificant because the 95% credible interval (approximated by bootstrapping) crosses the zero line.

[12] If manually labeling all the posts is the objective of the labeling tasks, then researchers need not be concerned about the performance of machine classifiers. If the researcher has the budget to build a full multimodal classifier, then they may build it and compare it with the text-only classifier—such a scenario is out of the scope of this article.

https://doi.org/10.1017/psrm.2025.10010 Published online by Cambridge University Press

recommendation is that researchers still obtain a subset of units labeled in a multimodal setting in order to evaluate the performance of the text-based classifier, as demonstrated in our example.

5. Conclusion

Researchers who conduct machine-assisted content analysis face an important trade-off regarding whether to utilize multimodal information in the process of labeling. While incorporating multimodal information carries additional costs in terms of time taken to label and reduced performance of text-based machine classifiers, it should improve the accuracy of the human annotated training data as evidenced by reduced proportion of invalid labels and improvements in intercoder reliability. This paper proposes a study design that can aid the researcher's decision on whether to incorporate multimodal information into the labeling process. We propose a set of performance metrics regarding the costs and benefits of multimodal labeling and an experimental framework that can help researchers estimate them in the pilot stage of their content analysis and illustrate the design with a tweet-labeling task. We show that failing to provide labelers with full multimodal content degrades the quality of labels returned, which is of first-order importance for applied researchers attempting to measure a quantity of theoretical interest. And we show that the penalty to downstream text-based classifier performance when using multimodal content for labeling highlights the need for applied researchers to pursue recent innovations in machine learning algorithms that can exploit both the text and non-text features of the training data.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psrm.2025.10010. To obtain replication material for this article, https://doi.org/10.7910/DVN/E2BV85.

Author contribution. HC, JB, and JN designed the research. HC built the tool for data collection. HC analyzed data. HC wrote the paper. All authors contributed to revisions.

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

Barberá P, Boydstun AE, Linn S, McMahon R and Nagler J (2021) Automated text classification of news articles: A practical guide. *Political Analysis* **29**, 19–42.

Barberá P, Casas A, Nagler J, Egan PJ, Bonneau R, Jost JT and Tucker JA (2019) Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review* **113**, 883–901.

Benoit K, Conway D, Lauderdale BE, Laver M and Mikhaylov S (2016) Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review* **110**, 278–295.

Gilardi F, Alizadeh M and Kubli M (2023) ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* **120**, e2305016120.

Grimmer J and Stewart BM (2013) Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* **21**, 267–297.

Hameleers M, Powell TE, Meer TGLAVD and Bos L (2020) A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication* **37**, 281–301.

Justin G, Roberts ME and Stewart BM (2021) Machine learning for social science: An agnostic approach. *Annual Review of Political Science* **24**, 395–419.

https://doi.org/10.1017/psrm.2025.10010 Published online by Cambridge University Press

King G, Lam P and Roberts ME (2017) Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science* **61**, 971–988.

Liu Y, Ott M, Goyal N, Jingfei D, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L and Stoyanov V (2019) RoBERTa: A robustly optimized BERT pretraining approach.

Peng Y, Lock I and Salah AA (2024) Automated visual analysis for the study of social media effects: Opportunities, approaches, and challenges. *Communication Methods and Measures* **18**, 163–185.

Peng Y, Lu Y and Shen C (2023) An agenda for studying credibility perceptions of visual misinformation. *Political Communication* **40**, 225–237.

Wu PY and Mebane WR (2022) MARMOT: A deep learning framework for constructing multimodal representations for vision-and-language tasks. *Computational Communication Research* **4**, 275–322.

Yingdan L and Shen CC (2023) Unpacking multimodal fact-checking: Features and engagement of fact-checking videos on Chinese TikTok (Douyin) *Social Media + Society* **9**, 205630512211504.

Zhang H and Pan J (2019) CASM: A deep-learning approach for identifying collective action events with text and image data from social media *Sociological Methodology* **49**, 1–57.

https://doi.org/10.1017/psrm.2025.10010 Published online by Cambridge University Press

Cite this article: Chen H, Bisbee J, Tucker JA and Nagler J (2026) Labeling social media posts: does showing coders multimodal content produce better human annotation, and a better machine classifier? *Political Science Research and Methods* **14**, 1019–1031. https://doi.org/10.1017/psrm.2025.10010