The Market for Information

Julien Berman

Presented to the Department of Economics
in partial fulfillment of the requirements
for a Bachelor of Arts degree with Honors

Harvard College
Cambridge, Massachusetts
May 25, 2026

Abstract

How does the availability of trustworthy news affect the production of online misinformation? Using a dataset of 2.7 billion Twitter posts and a panel of $9,833$ local newspapers, I show that increasing exposure to trustworthy local journalism significantly reduces the share of Twitter discourse that contains misinformation. I use text embeddings computed by a large language model to train supervised classifiers that predict whether tweets about four topic areas — vaccines, climate change, immigration, and U.S. election results — contain misinformation. In the two years after a newspaper changes its publication frequency, the share of vaccine-related tweets containing misinformation increases by 2.3 percentage points (15 percent of the pre-treatment mean) and the share of immigration-related tweets containing misinformation increases by 3.8 percentage points (8 percent of the pre-treatment mean). These effects are more pronounced for larger changes in coverage (e.g. the closure of a daily newspaper) than for smaller changes in coverage (e.g. a newspaper transitioning from publishing four days per week to publishing three days per week). Although Twitter users in Republican counties produce more misinformation than users in Democratic counties, treatment effects in Democratic counties are approximately twice the size of treatment effects in Republican counties for both vaccine-related misinformation and immigration-related misinformation.

Acknowledgements

Enormous thanks to Jesse Shapiro, Jim Snyder, Andrei Shleifer, and Aakaash Rao, without whom this thesis would not have been possible.

Contents

1 Introduction

The harms of misinformation are well-documented. For example, recent evidence suggests that false public health narratives about the Covid-19 pandemic contributed to vaccine hesitancy and impeded effective crisis response.[1] Social media platforms can amplify these harms. Because users can frictionlessly share content, and because social media feeds are determined algorithmically, misinformation on social media has the potential to spread rapidly to thousands or millions of users.[2]

Social media companies, government regulators, and academic researchers have explored many different policies to help curb the spread of misinformation. Most of these policies consist of some kind of "content moderation," in which social media platforms take down content, either manually or algorithmically, according to some pre-determined metric that evaluates whether speech is "false," "harmful," or otherwise worthy of removal.[3] Content moderation policies, however, require platforms and regulators to make contentious political choices about the types of speech that are permissible and the types of speech that are not.

In contrast, less research focuses on the structural features of the information ecosystem that lead people to produce and consume misinformation to begin with. This paper investigates one such structural question. In particular, I ask whether the production of misinformation depends on the availability of *trustworthy* information. Intuitively, trustworthy information and misinformation could either be complementary goods or substitute goods. Suppose that traditional news outlets with rigorous fact-checking and editing procedures primarily produce content that is less likely to contain false information, whereas the news available on social media is much more likely to contain false information. Then, trustworthy information and misinformation are *complements* if a greater demand for news increases consumption of both trustworthy news and misinformation. In contrast, trustworthy information and misinformation are *substitutes* if people allocate a fixed attention budget for news and choose what to read, watch, and share from a set of available sources. In this case, readily available trustworthy news from traditional media

[1] See, e.g., Bursztyn et al. (2023); Pennycook et al. (2021); Pennycook et al. (2020).

[2] See, e.g., Vosoughi et al. (2018).

[3] See Kominers and Shapiro (2025), Hossain et al. (2024), and Yang et al. (2023) for a discussion of content moderation policies.

sources can satisfy consumers' demand for information before they turn to less-trustworthy sources on social media.

Consequently, it is of importance to policymakers to understand the relationship between trustworthy information and misinformation. If an abundant supply of trustworthy information can "crowd out" the production of misinformation on social media, then policies that increase users' exposure to trustworthy news sources (or policies that incentivize users to receive information from trustworthy news sources) can reduce the spread of online misinformation before it takes hold. Indeed, if trustworthy news acts as a structural counterweight to false content, these policies can significantly strengthen the information ecosystem.

This paper studies whether trustworthy information can in fact operate as a substitute for misinformation. I use local newspaper coverage in each U.S. county as a proxy for the supply of trustworthy information available in a particular information market. I then examine the *volume* (amount of overall activity) and *composition* (share of overall activity that consists of misinformation) of discourse on Twitter produced by users located in each information market. Focusing on the period from January 2012 to May 2023, I leverage changes in local newspaper coverage during that time to test how the production of misinformation on Twitter responds to changes in the supply of trustworthy news.

Local newspaper coverage is a natural choice to measure the supply of trustworthy news. Unlike social media posts, newspaper articles must survive multiple layers of editorial review before publication, and journalists operate under professional codes of ethics that place a premium on factual accuracy (Society of Professional Journalists 2014). Readers recognize this distinction: Recent surveys conducted by Pew Research find that Americans view local news media as roughly twice as credible as social media (Pew Research Center 2025). Of course, local newspapers are not entirely free of misinformation, but the density of misinformation in traditional media is almost certainly far lower than on social media.

Over the past two decades, however, hundreds of local newspapers have closed or reduced publication frequency, sharply reducing the availability of local reporting in many counties. I assemble a novel panel dataset that tracks newspaper entries, exits, and changes in publication frequency (hereafter collectively referred to

as "coverage changes") for each U.S. county and each month during the sample period. In total, I track $9,833$ unique newspapers. For 448 of these newspapers, I also obtain the number of articles published each month by scraping NewsLibrary.com, which is an online news database that houses a conglomeration of local news. To my knowledge, this panel is the most detailed and granular dataset ever created for U.S. newspaper markets during the sample period.[4] The dataset describes the supply of trustworthy news provided by local newspapers to their readers.

To measure Twitter discourse in each information market, I use the Geotweet Archive maintained by the Center for Geographic Analysis (CGA) at Harvard University, which contains 2.7 billion geotagged Twitter posts ("tweets") posted during the sample period by users located in the United States. I sort these tweets into four topic areas I have identified as misinformation-prone: vaccines, climate change, immigration, and U.S. election results. Then, for each of these four topic areas, I compile separate datasets consisting of tweets about the relevant topic area that have already been labeled by researchers as either likely to be true or likely to be false.

Using the large-scale Geotweet Archive as well as the separate labeled datasets, I introduce and test a new methodology to detect misinformation on social media. For each topic area, I use the labeled datasets to train supervised machine-learning classifiers that predict whether a tweet contains false information based on the text of the tweet. I show that text "embeddings"[5] computed using a state-of-the-art large language model contain extraordinary predictive power for determining whether a tweet contains false information. Across all classifiers, the lowest out-of-sample AUC was 0.829, and the majority of classifiers had AUCs above 0.92. Then, I apply each supervised classifier at scale and predict the extent to which each relevant tweet in the Geotweet Archive is likely to contain misinformation.

This approach for identifying misinformation on social media is different from the approach traditionally used in the economics literature to study misinforma-

[4] Ewens et al. (2022) compiles a similar panel from 2001 to 2017. However, our datasets differ in two key respects. First, Ewens et al. (2022) only considers daily newspapers in the sample. Second, due to certain data limitations, the researchers organize their panel at a yearly level. My panel, in contrast, includes both daily and weekly newspapers, and it records precise dates for all coverage changes.

[5] A text embedding is a high-dimensional vector representation of a string of text.

tion. For example, Allcott and Gentzkow (2017) measures the spread of 156 entirely fabricated news articles that were subsequently shared on Facebook. Similarly, Vosoughi et al. (2018) tracks the proliferation of $126,000$ rumors on Twitter. (The authors identify a "rumor" as any tweet that contains a claim that is explicitly debunked by a verified fact-checking website.[6]) On the one hand, the approaches taken by Allcott and Gentzkow (2017) and Vosoughi et al. (2018) have the advantage that the content they classify as misinformation has been verified as false by professional fact-checkers, minimizing the likelihood of misclassification. On the other hand, both approaches for identifying misinformation rely on predefined lists of fact-checked stories, so they likely miss a substantial share of misinformation on social media that is never reviewed by these organizations. In contrast, my approach predicts the veracity of every tweet that pertains to one of the four misinformation-prone topic areas. As a result, I pick up content containing misinformation that has not yet been manually debunked by a professional fact-checker. However, because my classifiers are imperfect, these predictions are subject to classification error. This tradeoff is unavoidable, though, if I aim to obtain broad coverage of potentially false content.

I provide a set of novel descriptive facts about the volume, composition, and geographic distribution of misinformation on Twitter pertaining to these four topic areas. In particular, I find that 24 percent of tweets about vaccines, 4 percent of tweets about climate change, 43 percent of tweets about immigration, and 71 percent of tweets about U.S. election results are likely to contain false content. In addition, I find that the prevalence of misinformation is strongly positively correlated across topics. For example, counties with a high share of false climate-related tweets also tend to have high shares of false immigration-related tweets and election-related tweets. Notably, there is a strong positive relationship for all topic pairs (β between 0.3 and 0.5), except for those that involve the share of false vaccine-related tweets (β between 0.04 and 0.07), suggesting that the geographic distribution of vaccine-related misinformation is somewhat distinct from the geographic distribution of misinformation about climate change, immigration, and U.S. election results. Finally, I find that users in counties with a greater Republican vote share in the previ-

[6]Specifically, the researchers focus on the following six websites: snopes.com, politifact.com, factcheck.org, truthorfiction.com, hoax-slayer.com, and urbanlegends.about.com.

ous presidential election are more likely to post tweets containing misinformation, suggesting that the prevalence of misinformation is positively correlated with partisanship. This relationship is persistent across all four topic areas (β between 0.6 and 1.7).

Next, I examine how discourse on Twitter responds to changes in the supply of local newspapers. Specifically, I estimate whether coverage changes affect (i) the total number of tweets posted by users in treated counties about each of the four topic areas, and (ii) the share of those tweets that are likely to contain misinformation. To estimate average treatment effects on ever-treated counties — counties that experience coverage changes at some point during the sample period — I exploit the fact that these coverage changes are staggered over time. Then, using an event-study difference-in-differences framework, I compare changes in Twitter discourse in ever-treated counties to changes in Twitter discourse in never-treated counties that do not experience coverage changes at any point during the sample period. This approach allows me to estimate the dynamic effects of a coverage change in the months and years following treatment.

My event-study difference-in-differences design rests on a standard set of identification assumptions developed by Callaway and Sant'Anna (2021) and Callaway et al. (2024) for settings with varying treatment timings and dosages. Most importantly, it requires that — absent a change in local newspaper coverage — ever-treated and never-treated counties would have followed "parallel trends" in the relevant Twitter outcomes. The design also assumes that counties do not systematically change their Twitter behavior in anticipation of future coverage changes, and that changes to local newspaper coverage in one county do not spill over into neighboring counties. I take several steps to ensure that this methodology is plausible and to probe the empirical implications of these assumptions. First, I conduct a simulation exercise to verify that I recover zero treatment effects using a placebo outcome variable designed to impose the sharp null of no causal treatment effects. Second, I test for pre-trends by examining whether treated and control counties exhibit differential trajectories prior to coverage changes. If there are no pre-trends, I fail to reject a key observable implication of the parallel trends assumption.

Ultimately, I find that coverage changes of local newspapers have no effect on

the *volume* of overall Twitter activity, but have large effects on the *composition* of discourse on Twitter, for certain topic areas. To determine treatment effects on the volume of Twitter activity, I estimate the event-study difference-in-differences design using, as the outcome variables, the total number of tweets posted in each county in each month about each of the relevant topic areas. The results of these event studies indicate that coverage changes affect tweet volumes by precisely zero percent for all topic areas. For all event studies, treatment effects in each post-treatment month hover around zero, and confidence intervals generally extend only a few hundredths of a percentage point in the positive or negative direction.

To determine treatment effects on the composition of discourse on Twitter, I estimate the event-study difference-in-differences design using, as the outcome variables, the share of tweets about the relevant topic that contain misinformation. The average coverage change roughly corresponds to 1.5 weekly newspapers going out of business. I estimate that, in the two years following a coverage change, the share of tweets about vaccines that contain misinformation increases by 2.3 percentage points (15 percent of the pre-treatment mean). Likewise, the share of tweets about immigration that contain misinformation increases by 3.8 percentage points (8 percent of the pre-treatment mean). I do not detect robust effects on the share of tweets about climate change that contain misinformation or the share of tweets about U.S. election results that contain misinformation.

I also leverage variation in treatment intensity to estimate how average treatment effects scale with the size of a coverage change. For vaccine-related misinformation and immigration-related misinformation, estimated treatment effects sizes are roughly proportional to treatment intensity. For example, I estimate that a newspaper decreasing its publication frequency by a single day increases the share of vaccine-related tweets that contain misinformation by about 1.6 percentage points, but that decreasing publication frequency by four days increases the share of vaccine-related tweets that contain misinformation by about 5.7 percentage points.

Finally, I find that coverage changes have different effects on the composition of Twitter discourse in Republican counties and Democratic counties. Although I show in my descriptive statistics that Republican counties are likely to have a

higher *baseline* share of misinformation, I estimate that coverage changes *increase* the share of tweets about vaccines that contain misinformation and the share of tweets about immigration that contain misinformation by approximately twice as much for Democratic counties as for Republican counties. These estimates suggest that Democratic counties are more elastic to changes in the supply of trustworthy news, perhaps because more users in Democratic counties received their information from local newspapers to begin with, or because counties with a lower baseline level of misinformation production have more room to shift in response to shocks.

These results support the hypothesis that trustworthy information and misinformation are substitutes and that trustworthy sources of news can crowd out misinformation. They are particularly significant given that the United States is experiencing a rapid contraction in local journalism, leaving many counties with limited sources of trustworthy news (see, e.g., Ewens et al. (2022)). If the loss of trustworthy news increases consumption and production of online misinformation, then the erosion of local journalism may generate large negative spillover effects on the American information ecosystem.

Still unknown, however, is the mechanism by which changes in local news coverage affect discourse on Twitter. Although I do not have substantial quantitative evidence, I propose two hypotheses about the behavior that could underlie the treatment effects I observe.

First, a decrease in local news coverage could induce people who were previously not on Twitter to join Twitter. Perhaps they previously read the local newspaper — either online or in print — but when the paper closed (or decreased coverage), they turned to social media instead. As a result, these users may be more exposed to content containing misinformation. If this were the mechanism driving the increase in the share of false tweets, I would observe large treatment effects on the *first-time users* posting tweets about each relevant topic area in the months following treatment, but lesser effects on the users who were already on Twitter.

In principle, this is testable in the data, because I observe unique user identifiers associated with each tweet. In practice, however, tweets about each of the relevant topic areas are relatively sparse, which means that the vast majority of users in a given county are first-time posters on the topic area of interest. As a result, I do not

have enough statistical power to estimate distinct treatment effects for first-time posters relative to veteran posters.

Second, a decrease in local news coverage could affect users that consume news both from their local newspaper and from Twitter. If these users receive less trustworthy information from the newspaper, they may be less equipped to dispel the rumors they encounter on social media. This mechanism is not testable with the data that I observe, because I would need information about the behavior of individual Twitter users' news diets.

This paper proceeds as follows. In section 2, I discuss related strands of literature and outline my contributions to each strand. In section 3, I explain the datasets used in my analysis, as well as the processes used to assemble them. In particular, I describe the methodology used to categorize tweets and determine which tweets contain misinformation. In section 4, I present descriptive facts about the spread of misinformation on Twitter over the course of the sample period. In section 5, I describe, in detail, the event study approach that I use to isolate causal effects of coverage changes, including identification assumptions and estimation procedures. In section 6, I present estimates for the causal effects of local newspaper coverage changes on Twitter discourse. I conclude in section 7.

2 Related literature

This paper contributes to three distinct strands of literature. First, a large body of work documents the diffusion of online misinformation (see Nyhan (2020) for a survey of the relevant literature). Most notably, Allcott and Gentzkow (2017) investigates misinformation in the context of the 2016 U.S. presidential election and finds that false news stories favoring Donald Trump were shared nearly four times as often as those favoring Hillary Clinton. Similarly, Vosoughi et al. (2018) shows that false news spreads farther and faster than true news on Twitter. In addition, several papers have studied policies implemented by social media platforms that can limit the spread of misinformation. For example, Allcott et al. (2019) and Chiou and Tucker (2018) show that Facebook's fact-checking and advertisement protocols successfully reduced user interactions with fake news sites and decreased the num-

ber of times anti-vaccine posts were shared. Ahmad et al. (2024) studies the companies that advertise on websites known to promulgate misinformation and finds that information-based platform interventions can reduce the financial incentives sustaining those fake news websites. Other work estimates the consequences of misinformation exposure on real outcomes: Bursztyn et al. (2023) finds that areas exposed to cable news narratives downplaying the threat of the Covid-19 pandemic had a greater number of Covid-19 cases and deaths, whereas Müller and Schwarz (2023) and Cao et al. (2023) connect Trump's xenophobic political rhetoric on Twitter to spikes in racially motivated hate crimes.

In contrast to the above empirical literature, which measures the *effects* of misinformation and explores policy interventions that can limit its spread, my paper instead investigates the *origins* of misinformation. Rather than asking how false content diffuses once it has emerged, I instead investigate the structural conditions that give rise to it in the first place, and I hypothesize that trustworthy sources of news can "crowd out" misinformation. There are several papers that provide theoretical frameworks for this hypothesis,[7] but none, to my knowledge, that test it in the data.

This paper is closely related to Campante et al. (2025), which uses a field experiment to show that increasing the salience of AI-generated misinformation reduces overall trust in news and increases demand for a news outlet that is already known to be trustworthy. However, whereas Campante et al. (2025) measures how the consumption of trustworthy news responds to changes in the supply of misinformation, I investigate how the consumption of misinformation responds to changes in the supply of trustworthy news.

Consequently, this paper also contributes to a nascent body of psychology literature that studies "prebunking" — an umbrella term to describe efforts to build resistance to misinformation before people encounter it (van der Linden et al. 2026; van der Linden 2024; Compton et al. 2021). Researchers typically study the efficacy of "prebunking" in a laboratory setting. Recently, though, Google has started deploying short prebunking videos that describe common misinformation tactics

[7]See, e.g., Acemoglu et al. (2024), which models platform network design and demonstrates that engagement incentives on social media can generate homophilic "filter bubbles" for low-reliability content.

alongside political content (Harjani et al. 2022). My findings suggest that trustworthy news sources could perhaps help increase resistance to misinformation in a similar way.

A second strand of literature studies the effects of newspapers on civic outcomes. Using a panel of newspaper entries and exits from 1869–2004, Gentzkow et al. (2011) shows that newspapers increase political participation, but have limited effects on partisanship. Snyder et al. (2010) uses the "congruence" between newspaper media markets and congressional districts to demonstrate that more newspaper coverage increases voter knowledge and improves accountability for elected officials. More recent research studies the changing industrial organization of local news: Ewens et al. (2022) shows that private equity buyouts affect newsroom staffing and content and lead to declines in civic engagement, and Heese et al. (2022) demonstrates that newspaper closures increase corporate violations and penalties.[8]

This literature, however, does not examine the effects of local newspapers on the broader online information ecosystem. In particular, although existing work traces how local news supply affects political participation, accountability of elected officials, and firm behavior, it does not test whether changes in the supply of local news affect the volume or composition of the information that is produced and circulated on social media.

Third, a robust literature in computer science explores the methods for algorithmically detecting online misinformation. Early research attempted to identify false content on social media using models exclusively trained on hand-coded characteristics about the text, author, and pattern of diffusion of each post (Castillo et al. 2011). More recent research uses recurrent neural networks (Ma et al. 2016), graph neural networks (Xu et al. 2022), and pre-trained transformers (Kaliyar et al. 2021) to classify misinformation without the need for manual feature engineering (see Alnabhan and Branco (2024) for a survey of the relevant literature). These modern approaches leverage large corpora of high-quality labeled datasets[9] used for

[8]See also Chiou and Tucker (2017), which finds that when Google News removed content published by the Associated Press, subsequent visits to news sites substantially decreased, and Athey et al. (2021), which leverages the 2014 shutdown of Google News in Spain to show that treated users reduced their overall news consumption.

[9]See Thibault et al. (2025) for a comprehensive discussion of available datasets.

training data-hungry models.

This paper contributes to this third strand of literature in two ways. First, to my knowledge, this is the first paper that attempts to identify misinformation at scale, for billions of social media posts. Previous studies typically apply models only within the datasets used to develop them. Second, although previous literature applies supervised machine-learning techniques to identify misinformation, this paper is the first to train classifiers using text embeddings computed by state-of-the-art large language models. I show that supervised machine-learning techniques are adept at detecting misinformation at scale when high-quality labeled data is easily accessible.

3 Data

My analysis draws on five types of data. This section describes the first three: (i) 2.7 billion individual Twitter posts geotagged by location of origin; (ii) issue-specific "ground truth" datasets consisting of pre-labeled true and false claims; and (iii) a panel of newspaper coverage changes and article output at the county-month level. The remaining two datasets are more standard: (iv) U.S. electoral outcomes from the CQ Voting and Elections Collection (CQ Press 2024); and (v) county-level demographic data from the U.S. Census Bureau (U.S. Census Bureau 2025).

3.1 Geotweet Archive

The Twitter data comes from the Geotweet Archive maintained by the Center for Geographic Analysis (CGA) at Harvard University (Lewis and Jain 2016). The Geotweet Archive is a global record of every geotagged tweet posted between January 2012 and May 2023, when Twitter closed access to its free API. Every tweet in the Geotweet Archive contains one of two types of geographical signatures. The first is a GPS-based latitude / longitude pair generated by the device posting the tweet. The second is a latitude / longitude pair that identifies the centroid of a user-defined place of origin. Typically, the latter designations are the names of towns or cities. According to the CGA, all tweets that possess one or both of these signatures are included in the Geotweet Archive, which is estimated to be approximately be-

tween 1 and 2 percent of all tweets posted during the sample period. In total, the Geotweet Archive consists of approximately 10 billion tweets, of which 2.7 billion are in English and originated in the United States.

For each tweet, I observe the latitude and longitude of the device making the post, the associated user, the text of the post, the date posted, and the number of likes and reposts received. I do not, however, observe any information about a tweet's view count, nor do I observe geotagged data for a user's likes or reposts. In other words, if a user in Phoenix, Arizona posts a tweet about immigration, I observe the number of likes it receives, but I do not observe the users who liked it.

The tweets are processed in three phases. First, I use shapefiles from the U.S. Census Bureau to sort each tweet by county and remove tweets that do not originate from the United States. I also perform standard text preparation for natural language processing, such as removing uninformative stop words (e.g. "and," "in," "the") and stripping punctuation and hyperlinks.

Next, I use a state-of-the-art open-source large language model — the Qwen3 eight billion-parameter mixture of experts model developed by Alibaba — to classify each tweet by topic area. I prompt the model to identify tweets about the four topics I have identified as misinformation-prone — vaccines, climate change, immigration, and U.S. election results — as well as tweets about sports or music. The latter two categories serve as "placebo" topic areas: Because online discourse about sports and music are unlikely to be affected by local newspaper closures,[10] these topic areas provide falsification tests for causal effects on the volume of Twitter activity in the misinformation-prone topic areas. Each tweet can be assigned multiple topics. The full classification schema is provided in appendix A.1.

Finally, I use a different state-of-the-art large language model — the Qwen3 eight billion-parameter embedding model — to represent the text of each tweet in a 4096-dimensional vector space. I normalize each vector to have a magnitude of one. These "embeddings" capture the semantic meaning of each tweet. They serve as the inputs for inference performed by supervised classifiers, which are trained on labeled "ground truth" datasets and described in section 3.2 below.

Figure 1 presents monthly tweet volumes overall and by topic area for the du-

[10] Although local newspapers often cover sports, the scope of the coverage tends to be limited to local sports teams. I particularly identify sports-related Twitter content about national sports.

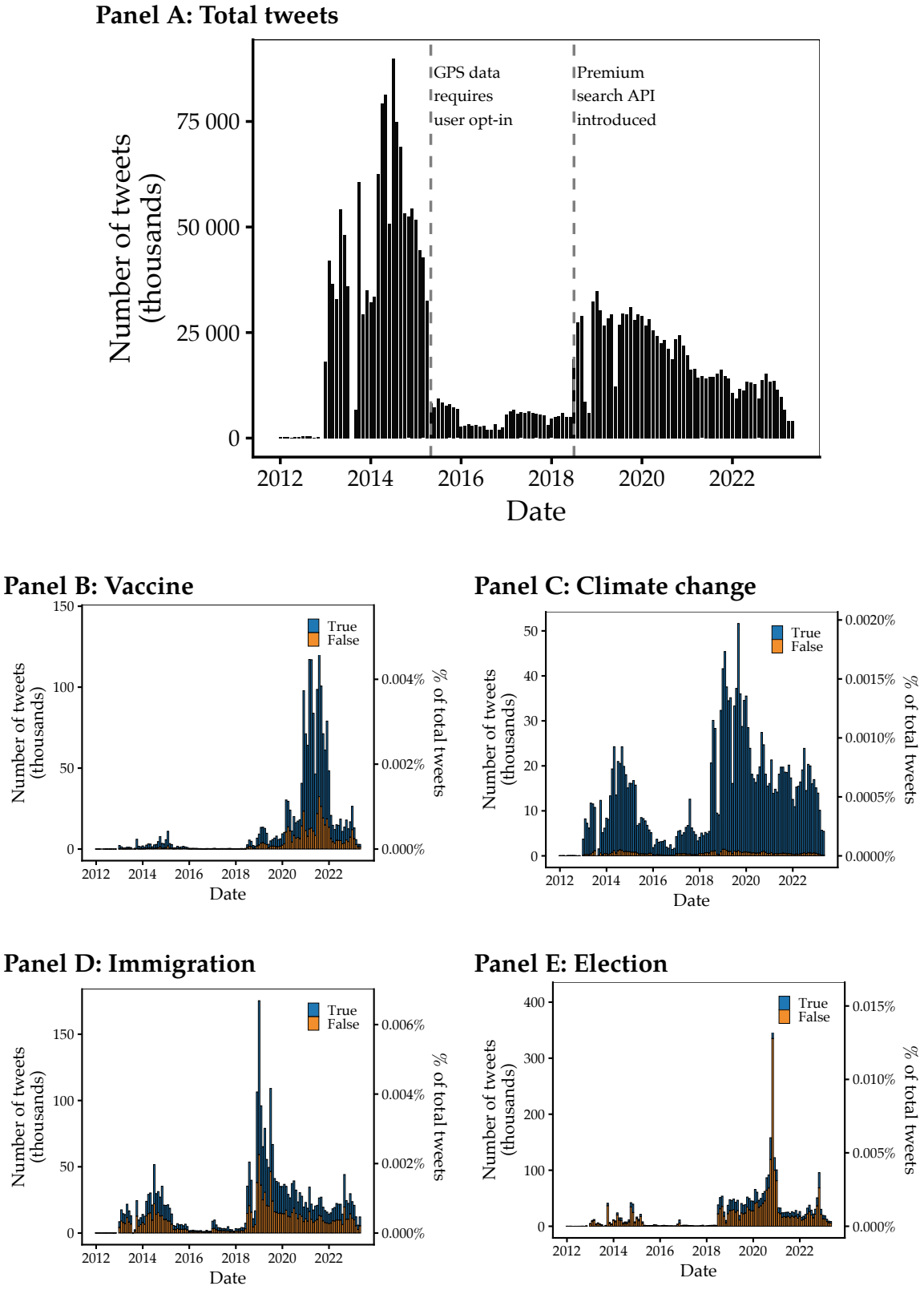
ration of the sample period. For each topic area panel, the left axis indicates the overall volume of tweets posted each month. The right axis indicates the volume of tweets posted each month as a percentage of the total number of tweets posted during the duration of the sample period.

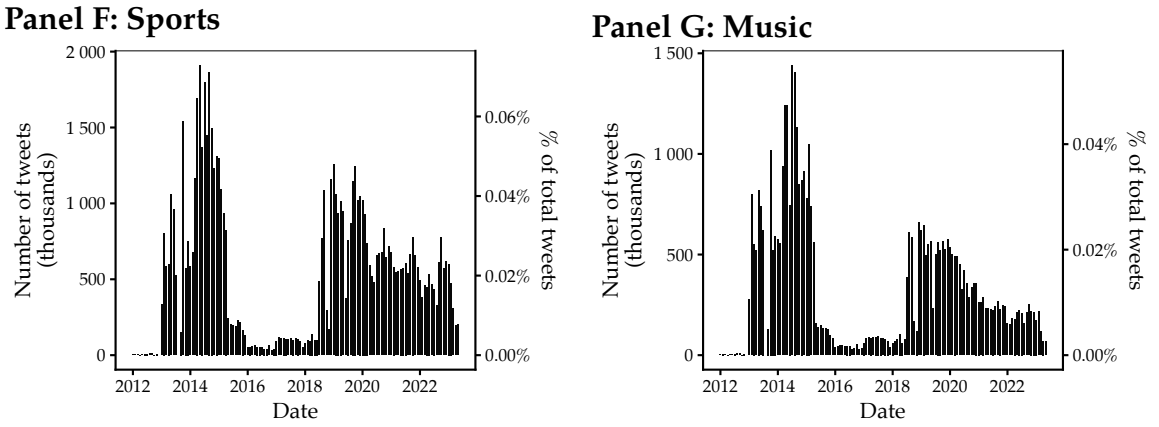
Notice that tweet volumes vary significantly by topic. For example, there are about ten times as many tweets about sports and music as there are about the next most common topic, immigration. In addition, the sample period can roughly be divided into three distinct "regimes." The first regime, from January 2012 to March 2015, contains the largest volume of tweets. Then, the number of posts per month falls abruptly in April 2015 before sharply increasing again in July 2018. These changes in monthly tweet volumes likely reflect changes in Twitter's policies. Before April 2015, GPS coordinate capture was enabled by default. As a result, many tweets contained geotags. After April 2015, users had to opt in to share their precise location (Cao et al. 2022). This policy change was partly counteracted in July 2018, when Twitter released a premium API feature that allowed for much more efficient scraping.

Beyond the idiosyncratic patterns in overall monthly tweet volumes, the topic-specific monthly tweet volumes match expected patterns. For example, the number of tweets about vaccines skyrocketed during the Covid-19 pandemic. The number of tweets about climate change is largest in 2019, when Greta Thundberg launched the global climate strike. The number of tweets about election results noticeably spikes around November on election years, particularly in 2020 and 2022.

Figure 2 provides a simple validation check for the Twitter data by examining whether county-level tweet volume scales sensibly with county population. In both panels, the dependent variable measures the natural log of the total number of tweets posted by users in a given county over the entire sample period. In panel A, the independent variable measures the natural log of a given county's population at the beginning of the sample. In panel B, the independent variable measures the natural log of a given county's population at the end of the sample. Because both axes are in logs, the slope of the relationship can be interpreted as an elasticity. For example, a 1 percent increase in a county's population at the beginning of the sample period is associated with a 1.23 percent increase in the number of tweets

Figure 1: Tweet volumes





Notes: This figure plots the volume of geotagged tweets posted during each month in the sample period. Panel A plots the total number of tweets posted each month. The dotted gray lines indicate the first months in which Twitter's two policy changes were implemented. Panels B-E plot the total number of tweets posted each month that pertain to each of the four topic areas I have identified as prone to misinformation. Panels F and G plot the total number of tweets posted each month that pertain to the two placebo topic areas that I have identified as likely to remain unaffected by changes in newspaper coverage. The left axis indicates the raw tweet volume. The right axis indicates the tweet volume as a percentage of the total number of tweets posted over the entire sample period. Blue bars indicate the number of tweets in each month that the relevant supervised classifier labels as likely to be entirely true. Orange bars indicate the number of tweets in each month that the relevant supervised classifier labels as likely to contain misinformation.

published by users within that county over the course of the sample period.

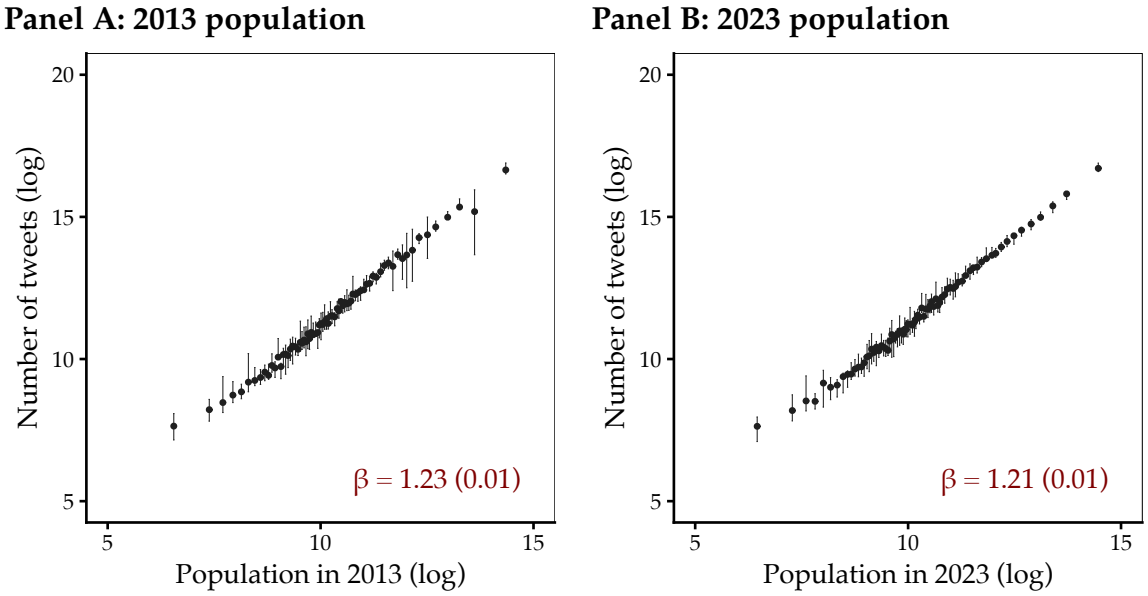
In both panels, the binscatter points lie close to a straight line with slope near one and narrow confidence intervals throughout the support, indicating a strong linear relationship between log population and log tweet volume. This pattern is reassuring; it suggests that the Geotweet Archive captures a geographically sensible measure of Twitter activity, in which tweet volume scales proportionally with the size of the underlying population.

3.2 "Ground truth" datasets

In order to determine which tweets contain misinformation, I rely on issue-specific "ground truth" datasets that consist of claims that have already been labeled as either "true" or "false." I consider these labels to provide a "ground truth" for whether a claim contains misinformation.

(1) **AntiVax.** The ground truth corpus for the "vaccine" topic area comes from

Figure 2: Relationship between number of tweets and population



Notes: This figure plots the relationship between a county's population and its Twitter activity as recorded in the Geotweet Archive. In both panels, the y-axis measures the natural log of the total number of tweets posted in a given county over the entire sample period. In panel A, the x-axis measures the natural log of the population of a given county at the start of the sample period. In panel B, the x-axis measures the natural log of the population of a given county at the end of the sample period. The univariate regression coefficient is reported in red. For each panel, I exclude from the sample all counties with fewer than 100 tweets over the entire sample period. Binscatters are estimated using the approach outlined in Cattaneo et al. (2024). The support of the independent variable is partitioned into quantile-spaced bins and, within each bin, the conditional mean of the dependent variable is computed. The number of bins is selected in a data-dependent manner. Error bars represent pointwise 95 percent confidence intervals. Standard errors are clustered at the county level.

the AntiVax dataset, which consists of approximately 15,000 English tweets posted between December 1, 2020 and July 31, 2021 that contain keywords related to vaccines and the Covid-19 pandemic (Hayawi et al. 2022). These tweets were hand-labeled by the researchers using information obtained from *Public Health*, the Centers for Disease Control and Prevention (CDC), and other official sources, and then verified by medical experts in public health.

(2) **CLIMATE-FEVER.** The ground truth corpus for the "climate change" topic area comes from the CLIMATE-FEVER dataset, which consists of approximately $1,500$ claims about climate change from online message boards sourced via Google search (Diggelmann et al. 2020). Climate scientists labeled each claim as either "true," "false," or "unknown." Evidence supporting and/or refuting each claim was sourced from Wikipedia and attached to the dataset.

(3) **TruthSeeker.** The ground truth corpora for the "immigration" and "election" topic areas come from the TruthSeeker dataset, which consists of approximately $180,000$ labeled tweets about a variety of topics (Dadkhah et al. 2023). To source the tweets, researchers first selected 700 true statements and 700 false statements from PolitiFact, a non-partisan fact-checking website operated by the Poynter Institute. Then, researchers manually generated keywords associated with each statement and crawled the Twitter API to find related tweets. Finally, 456 Amazon Mechanical Turk users labeled each tweet based on how strongly the claim in the tweet agrees with the PolitiFact statement. Based on these labels, each tweet was assigned to one of five categories: "agree," "mostly agree," "no majority," "mostly disagree", "disagree," and "unrelated." I remove all tweets in the "no majority" and "unrelated" categories, and, based on the veracity of the original PolitiFact statement, assign each tweet to be either "true" or "false." In order to identify the subset of tweets about immigration and U.S. elections, I prompt the same Qwen3 mixture-of-experts large language model with the same schema that I used to classify the Geotweet Archive. In total, I obtain 5,497 labeled tweets in the immigration topic area and 7,749 labeled tweets in the elections topic area from the TruthSeeker dataset.

For each of the above datasets, I preprocess the labeled claims exactly as I preprocessed the tweets in the Geotweet Archive. I remove stopwords and use the

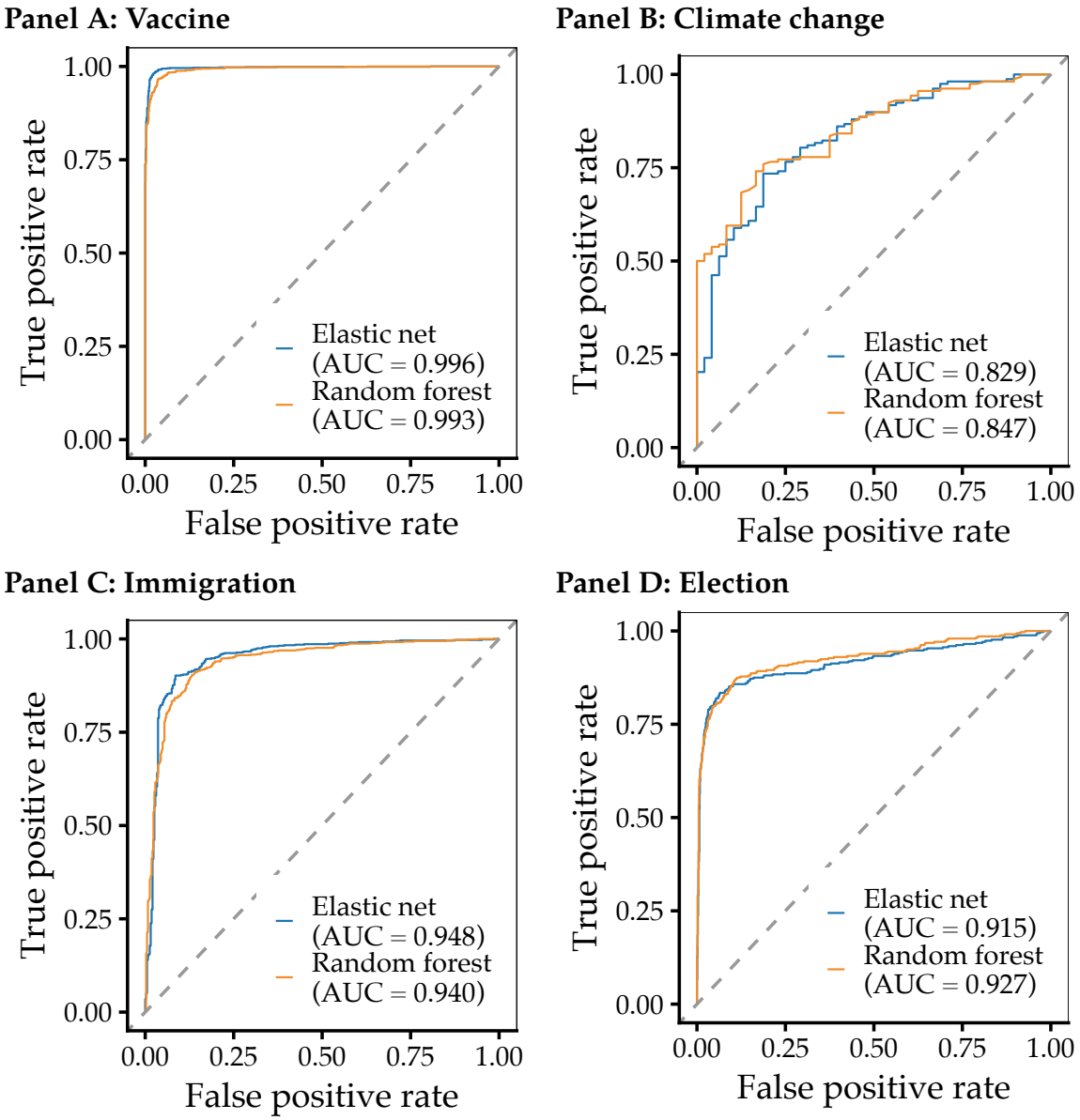
Qwen3 embedding model to obtain 4096-dimensional vector representations of each claim, which I then normalize to have a magnitude of one. Formally, let $D_i = \{(e_j, y_j)\}_{j=1}^{n_j}$ denote the labeled data for topic area i, where $e_j \in \mathbb{R}^{4096}$ is the embedding representation of claim j and $y_j \in \{0, 1\}$ indicates whether the claim is false. For each topic area i, I train two supervised models — an elastic net and a random forest — that take e_j as inputs and predict y_j. Appendix A.2 and A.3 describe each classifier in more detail.

Figure 3 presents receiver operating characteristic (ROC) curves for each model. The ROC curve plots the true positive rate against the false positive rate at all possible classification thresholds. Models that perform better bow toward the upper-left corner, achieving high true positive rates without incurring many false positives. Models that randomly guess the class of each tweet lie along the 45-degree diagonal. The area under the ROC curve (AUC) summarizes overall performance. A perfect classifier has an AUC of 1, whereas random guessing yields an AUC of 0.5. All ROCs and AUCs are computed on a holdout set that the classifier did not observe during model training.

Performance varies substantially across topic areas. For example, both the elastic net and random forest models trained on the AntiVax dataset achieve extremely high AUCs (0.996 and 0.993, respectively), indicating that the semantic content of vaccine misinformation differs markedly from that of true claims such that the two models are easily able to distinguish between the two. In contrast, both the elastic net and random forest models trained on the CLIMATE-FEVER dataset perform considerably worse (0.829 and 0.847, respectively). I hypothesize that this discrepancy reflects differences in the nature of the claims used in the training: Statements in the CLIMATE-FEVER dataset tend to be more technical and clinical, so true and false claims may not differ as sharply in their semantic content. In addition, I observe fewer labeled claims in the CLIMATE-FEVER dataset than I do for the other topic areas, which could also contribute to poor model performance.

Table 1 provides examples of two true claims and two false claims for each topic area. These claims are drawn randomly from the appropriate holdout set used to evaluate out-of-sample accuracy for each classifier. Column 3 lists the label assigned by the human coders, and column 4 lists the label predicted by the relevant

Figure 3: Model performance



Notes: This figure plots ROC curves for the elastic net and random forest models described in appendix A.2 and A.3. The area under the ROC curve (AUC) is provided for each model. All ROC curves and AUC statistics are calculated on a subset of the training dataset that is unseen by the model prior to evaluation. Each panel corresponds to one topic area. Training data for models summarized in panel A come from the AntiVax dataset. Training data for models summarized in panel B come from the CLIMATE-FEVER dataset. Training data for models summarized in panels C and D come from the TruthSeeker dataset. These datasets are described in section 3.2 of the text.

classifier.

3.3 Newspaper entries and exits

I construct a detailed panel of local newspaper coverage from multiple sources. First, the U.S. News Deserts Database Archive at the University of North Carolina provides cross-sectional snapshots of the local news ecosystem in the United States for the years 2004, 2010, 2012, 2014, 2016, and 2020 (Center for Innovation and Sustainability in Local Media 2024). However, these data have large holes. Many local newspapers, particularly smaller ones, are missing, and I do not observe each newspaper frequently enough to precisely pin down dates of entry and exit. Consequently, I supplement the News Deserts database with data provided by Jim Snyder, which fills many of the holes in the News Deserts database and contains more precise dates of coverage changes.

In addition, I hand-collect detailed information about the largest 291 newspapers that changed their coverage frequency at some point during the sample period. To do this, I contact the librarians who service the communities covered by the local papers. From the librarians, as well as news articles that mention changes to the local paper, I collect:

- The precise dates that the local newspaper changed coverage frequency;
- A dummy variable indicating whether the newspaper has an online presence;
- Details about the transition, including information about mergers, acquisitions, and layoffs;
- A list of covered municipalities; and
- A list of alternative local news sources that cover the county that the newspaper services.

In total, I observe $9,833$ distinct newspapers over the course of the sample period.

For 448 newspapers in my sample, I scrape NewsLibrary.com — an online news database that houses articles published online by many different local newspapers — to obtain the number of articles published by each newspaper in each month.

Table 1: Examples of labeled claims

Topic	Text	Label	Pred
Vaccines	india's covid vaccine shortage: the desperate wait gets longer	True	True
Vaccines	just over a week until i am 24 and get my covid vaccine... yay?	True	True
Vaccines	but what about the fact it is an experimental biological agent that will not prevent transmission of the virus? dr. simone gold is a truth-telling doctor. listen to her talk, "the stand". we are not sheeple. we are conscious and we are aware and we see.	False	False
Vaccines	ok what i wanna understand is how the woman who taught me that mainstream doctors do everything they can to rip you off so we must do our best to heal naturally is the same one who thinks its crazy that i dont wanna take the rushed vaccine being pushed by those same doctors	False	False
Climate change	however this is exactly what climate scientists have predicted for california since at least the 1980s protracted periods of warm dry conditions punctuated by intense wet spells with more rain and less snow causing both drought and floods	True	True
Climate change	the most notorious was 252 million years ago it began when carbon warmed the planet by five degrees accelerated when that warming triggered the release of methane in the arctic and ended with 97 percent of all life on earth dead	True	True
Climate change	sea level rise has been slow and a constant predating industrialization	False	False
Climate change	humanproduced carbon might be one of the factors of climate change but theres simply no evidence that it is a significant one	False	False

Topic	Text	Label	Pred
Immigration	obama/clinton elected by us organized crime groups that facilitated us taxpayer financed international crime w/foreign organized crime groups w/ anti-us religious/political ideologies ; the criminal activities include heroin/meth ; human trafficking through our southern border.	True	True
Immigration	: fact: mitt romney said hed veto the dream act and called for undocumented workers to self-deport.	True	True
Immigration	1) our southern borders are not open. stop repeating falsehoods. *the united states will temporarily limit inbound land border crossings from canada and mexico to essential travel. *this action does not prevent u.s. citizens from returning home. *these restrictions are	False	False
Immigration	what a bunch of crap! trumps handling of the border was to separate children from their families ; throw them in cages with no plan for then to rejoin the families. biden accomplished this task. he also solved covid, lowered unemployment ; is rebuilding the failed economy!	False	False
Election results	77,000 votes vote margin: pa = 44,292 wi = 22,748 mi = 10,704 less than one vote per precinct (and far less in wi and mi) (people in northern wisc and up - knowledge from my family there - were getting absolutely pounded on fb w/ propaganda in week or two before election.)	True	True
Election results	and, yet, black voter turnout was not there in 2016, like it was in 2008 and 2012.	True	True
Election results	sad to see how pathetic biden supporters are. cheering for a dictator that wants to raise tax, ship our jobs to china, inject deadly virus to win election and mandate a nationwide lockdown so he can lock us up in our own home prison. no wonder china is dare to takeover. weak usa!	False	False
Election results	the rate of rejected mail-in ballots is 30 xs lower in pennsylvania this year than it was in 2016. this is why they kept our poll watchers and observers out of the sacred vote counting rooms!	False	False

Notes: This table provides examples of labeled tweets for each topic area. Each topic area includes two true and two false claims randomly drawn from the relevant holdout set used to test out-of-sample accuracy. Column 1 provides the topic area. Column 2 provides the text of the tweet. Column 3 provides the label assigned by the human researchers. Column 4 provides the predicted label generated by the relevant topic-specific supervised classifier.

I collect article counts in order to verify that coverage changes do, in fact, reduce article output.

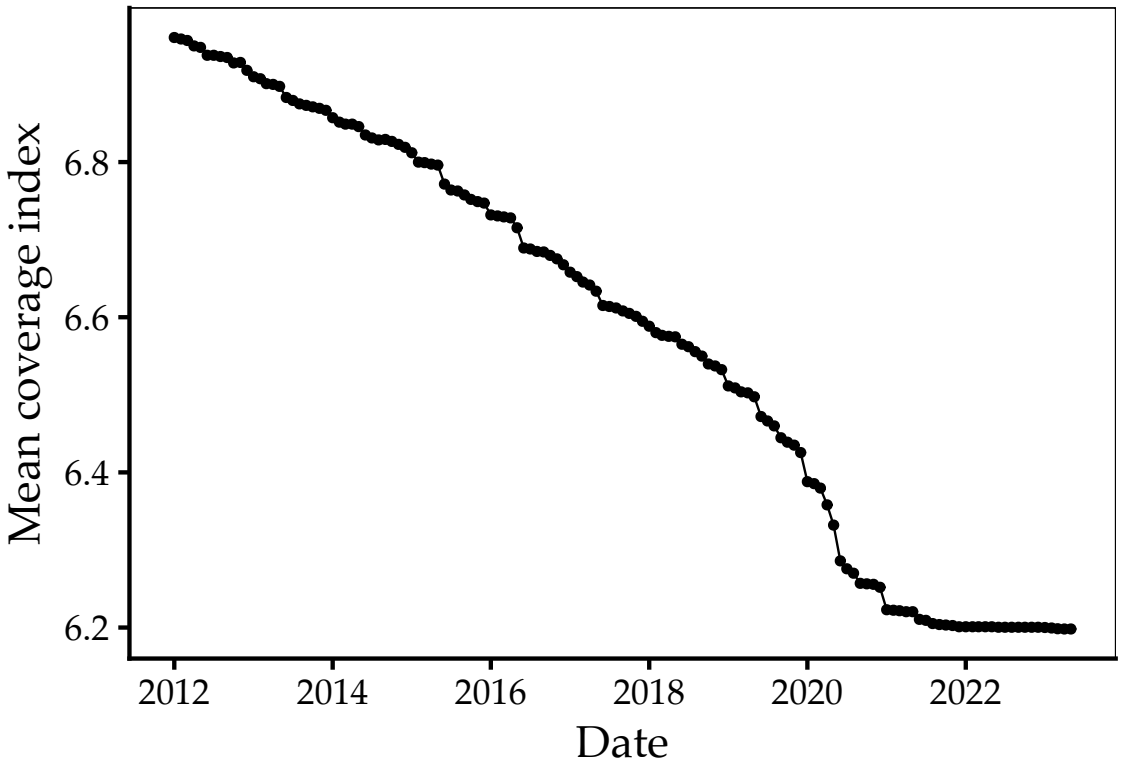
Using this panel of newspapers, I construct a "coverage index" for each newspaper. The coverage index measures, for each month, the number of days that the newspaper publishes. For many newspapers and months in both the News Deserts dataset and the data received from Jim Snyder, this measure is missing. Instead, I often only observe whether the newspaper is a "weekly" newspaper or a "daily" newspaper. Consequently, I use the 291 hand-collected newspapers to train an elastic net model to impute the coverage index for newspapers and months for which the coverage index is missing. I then compute the coverage index for a given county in a given month as the sum of the coverage indices for all its constituent newspapers. I define an "event" to be any instance in which the coverage index for a given county changes between two periods. More details about the construction of the coverage index are provided in appendix B.

The coverage index measures (roughly) the number of newspaper-days per week published in a given county in a given month. Consider Orleans County, Louisiana. From January 2012 to October 2012, Orleans County had one daily newspaper — the *Times-Picayune* — with a coverage index of 6 and three weekly newspapers with coverage indices of 2, for a total coverage index of 12. Then, in November 2012, the *Times-Picayune* cut down to publishing only 3 days per week, so the coverage index decreased from 12 to 9. In September 2013, a new daily newspaper — the *New Orleans Advocate* — opened, increasing the coverage index for Orleans County from 9 to 15. Orleans County therefore has two events, one in November 2012 and one in September 2013.

Figure 4 plots the mean coverage index for each month during the sample period. The coverage index steadily declines from about 7.0 in 2012 to about 6.2 in 2023 as many local newspapers decreased their publication frequency or closed altogether. A coverage index of 6 (roughly) corresponds to either a single daily newspaper or three weekly newspapers.

Figure 5 plots a timeline of all events (coverage changes) that take place during the sample period. The size of each rectangle measures the magnitude of the coverage change. There are many more negative coverage changes than there are positive

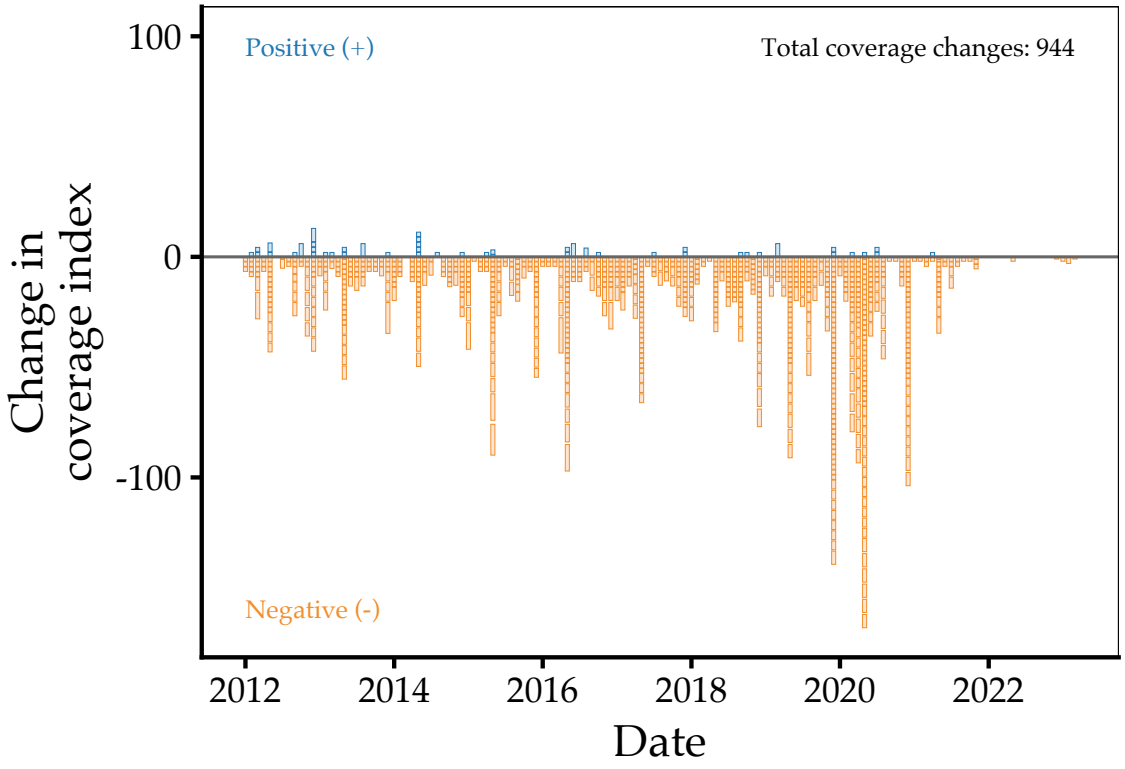
Figure 4: Newspaper coverage



Notes: This figure plots the mean coverage index, aggregated over all counties, for each month in the sample. The coverage index for a given newspaper can be interpreted as the estimated number of days-per-week that the newspaper publishes. The coverage for a given county is the sum of the coverage indices for each newspaper that services that county.

ones, confirming the shrinking size of the local newspaper industry. The periods with the most coverage changes overlap with the onset of the Covid-19 pandemic. In total, I observe 1100 coverage changes at the newspaper level, and 944 coverage changes at the county level.

Figure 5: Timeline of events



Notes: This figure plots the dosage of each coverage change at the county level for every month in the sample period. Events that add newspaper coverage (e.g. a newspaper entry) are plotted in blue. Events that reduce newspaper coverage (e.g. a newspaper exit) are plotted in orange. In total, 944 counties have coverage changes during the sample period.

4 Descriptive facts

I now establish several basic facts about the spread of misinformation on Twitter. I present county-level maps of the United States with the geographic distribution of misinformation by topic area. Then, I show that Republican counties have a larger share of tweets containing misinformation than Democratic counties, for nearly all

topic areas. Finally, I find that the share of tweets containing misinformation is positively correlated across topic areas.

4.1 Summary statistics

Table 2 reports summary statistics for the key variables used in my analysis. Each panel is organized based on the unit of observation. Panel A is at the county-month level. The "total tweets" variables measure the total number of tweets posted in a given county in a given month, by topic area. These variables are used to measure the overall volume of Twitter activity. The "share false" variables are computed by taking the total number of tweets that the relevant supervised classifer labels as likely to contain misinformation and dividing by the total number of tweets, for a given county in a given month, by topic area. These variables are used to measure the composition of Twitter discourse on a given topic. The coverage index is the key variable used to determine the intensity of treatment — the magnitude of a coverage change — when calculating the effects of exposure to local news on the composition of misinformation.

Because I construct my panel at the county-month level, many of the variables measuring Twitter activity are relatively sparse. Although the Geotweet Archive is extremely large in the aggregate, dividing the data across $3{,}144 \times 137 = 430{,}728$ observations produces many county-month cells with no tweets in a given topic area. Notice, for example, that the median county-month contains just 237 total tweets, and it contains zero tweets about vaccines, climate change, immigration, or U.S. elections.

Panel B is at the county level and summarizes treatment exposure. About 20 percent of counties experience at least one change in the newspaper coverage index during the sample period, and the distribution of the number of treatments is concentrated at zero with a long right tail. Some counties, though, experience as many as seven coverage changes over the course of the sample duration.

Panel C is at the newspaper level and summarizes newspaper characteristics used to construct county-level coverage. The statistics highlight that the typical paper in the underlying panel is weekly rather than daily, which is relevant for interpreting the discrete changes in the county coverage index as combinations of

Table 2: Summary statistics

	Mean	Std. dev.	Min	Median	Max
Panel A: County-month					
Total tweets	6,018.57	37,386.59	0.00	237.00	4,047,651.00
Total tweets: Vaccine	4.29	48.26	0.00	0.00	6,495.00
Total tweets: Climate change	4.20	29.98	0.00	0.00	3,011.00
Total tweets: Immigration	6.44	50.20	0.00	0.00	6,712.00
Total tweets: Election	7.52	75.93	0.00	0.00	18,760.00
Total tweets: Sports	165.35	960.11	0.00	5.00	95,697.00
Total tweets: Music	101.98	734.55	0.00	2.00	73,496.00
Share false tweets: Vaccine	0.23	0.30	0.00	0.10	1.00
Share false tweets: Climate change	0.03	0.13	0.00	0.00	1.00
Share false tweets: Immigration	0.49	0.35	0.00	0.50	1.00
Share false tweets: Election	0.68	0.32	0.00	0.73	1.00
Coverage index	6.58	10.02	0.00	4.00	244.00
Observations: 430,728					
Panel B: County					
Ever treated	0.21	0.41	0.00	0.00	1.00
Number of treatments	0.30	0.70	0.00	0.00	7.00
Observations: 3,144					
Panel C: Newspaper					
Coverage index	2.10	1.65	0.00	2.00	6.00
Daily newspaper	0.13	0.33	0.00	0.00	1.00
Weekly newspaper	0.87	0.33	0.00	1.00	1.00
Has NewsLibrary	0.05	0.21	0.00	0.00	1.00
Hand-collected	0.03	0.17	0.00	0.00	1.00
Observations: 9,833					
Panel D: Tweet					
Topic: Vaccine	0.00	0.03	0.00	0.00	1.00
Topic: Climate change	0.00	0.03	0.00	0.00	1.00
Topic: Immigration	0.00	0.03	0.00	0.00	1.00
Topic: Election	0.00	0.04	0.00	0.00	1.00
Topic: Sports	0.03	0.16	0.00	0.00	1.00
Topic: Music	0.02	0.13	0.00	0.00	1.00
Likes	0.64	269.34	0.00	0.00	1,427,982.00
Retweets	0.21	93.97	0.00	0.00	715,280.00
Observations: 2,621,250,705					

Notes: This table provides the mean, standard deviation, minimum, median, and maximum values for key variables observed in the data. Panel A provides summary statistics for variables observed at the county-month level. Panel B provides summary statistics for variables observed at the county level. Panel C provides summary statistics for variables observed at the newspaper level. Panel D provides summary statistics for variables observed at the tweet level.

entries and exits of weekly papers, entries and exits of daily papers, daily-to-weekly transitions, and other coverage adjustments.

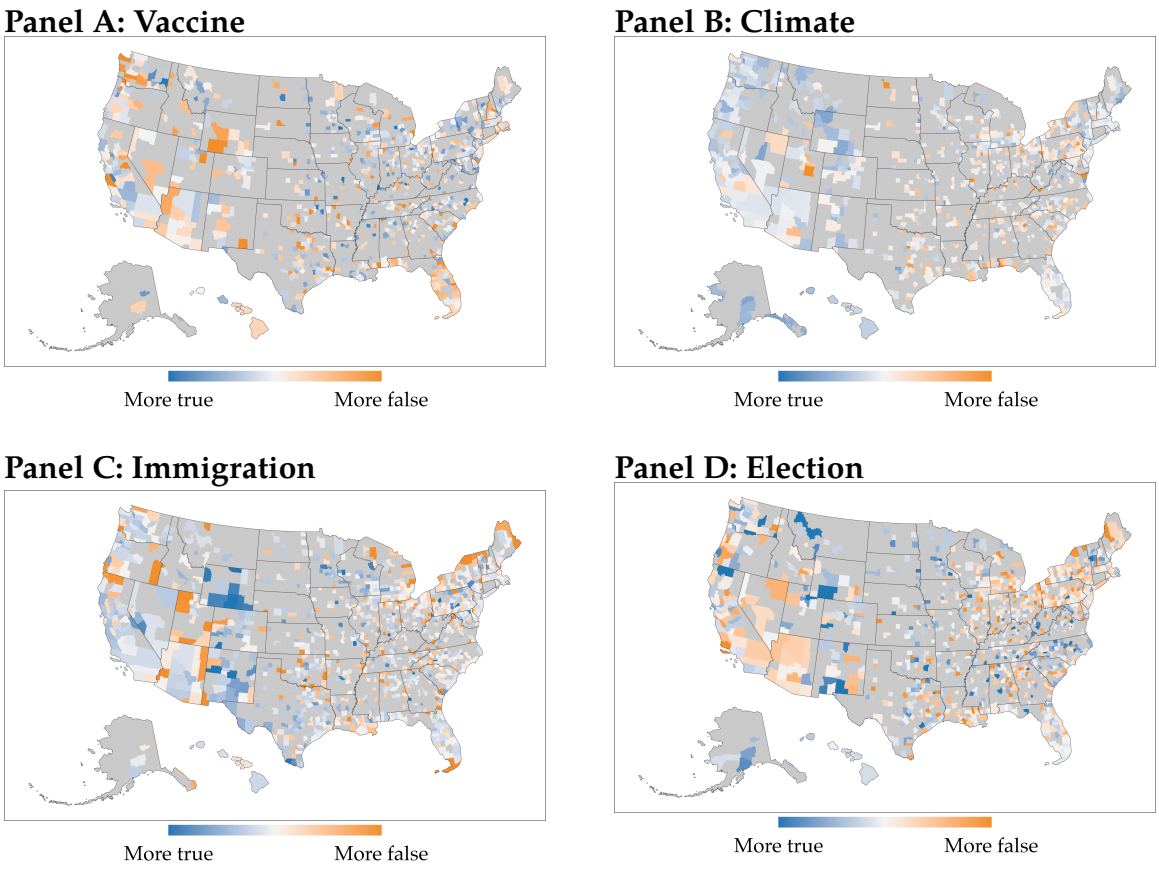
Panel D is at the tweet level and summarizes the raw Twitter variables. Two patterns are important. First, a relatively small share of the total Geotweet Archive pertains to each of the four topic areas I have identified as prone to misinformation. In fact, the mean of all four of the indicator variables that identify these topic labels is zero, up to two decimal points of precision. Second, the number of likes and retweets that each post receives is highly concentrated at zero with a heavy right tail. The median tweet receives zero likes and zero reposts, but maxima for both variables are extremely large.

4.2 Geographic distribution of misinformation

Figure 6 illustrates the geographic distribution of misinformation across counties for each of the four misinformation-prone topic areas. The underlying object in each panel is the share false measure: the share of topic-specific tweets that are predicted by the relevant supervised classifier to contain misinformation. To make differences comparable across topics with different baseline rates, I standardize this measure within each topic by subtracting the topic-specific mean and dividing by the topic-specific standard deviation. Counties shaded in warmer colors therefore have misinformation shares that are higher than average for that topic, and counties shaded in cooler colors have misinformation shares that are lower than average for that topic. To avoid over-interpreting noise from extremely small denominators, I exclude counties with fewer than 100 tweets in the relevant topic area over the full sample period.

Two broad patterns emerge. First, misinformation shares exhibit substantial spatial heterogeneity even after restricting to counties with meaningful topic-specific tweet volume: Some areas consistently sit well above the mean, while others sit well below it. Second, the spatial pattern is not identical across topic areas, consistent with the idea that misinformation is topic-dependent: Counties that are relatively misinformation-prone in one topic area need not be equally misinformation-prone in another. These descriptive patterns motivate the paper's focus on topic-specific outcomes and the use of within-topic composition measures in the causal analysis.

Figure 6: Misinformation maps



Notes: This figure plots the county-level geographic distribution of misinformation in each of the four topic areas I have identified as prone to misinformation. The underlying measure is the share of tweets in a given topic area that the relevant supervised classifier predicts contains misinformation. This variable is then standardized across counties by subtracting the mean and dividing by the standard deviation. Colors indicate how many standard deviations a county's share of misinformation for a particular topic area is above or below the sample mean for that topic area. For each panel, I exclude from the sample all counties with fewer than 100 tweets in the relevant topic area over the entire sample period.

Next, figure 7 documents a robust descriptive relationship between political partisanship and misinformation. Each observation is a county-election cycle, where an election cycle is defined as the four-year window from December of a presidential election year through November of the next presidential election year. Within each county-cycle, I aggregate all tweets in a given topic and compute the share false measure. As in the maps, I standardize the share false measure within each topic, and I exclude county-cycles with fewer than 100 topic-specific tweets. I report univariate regression coefficients in red.

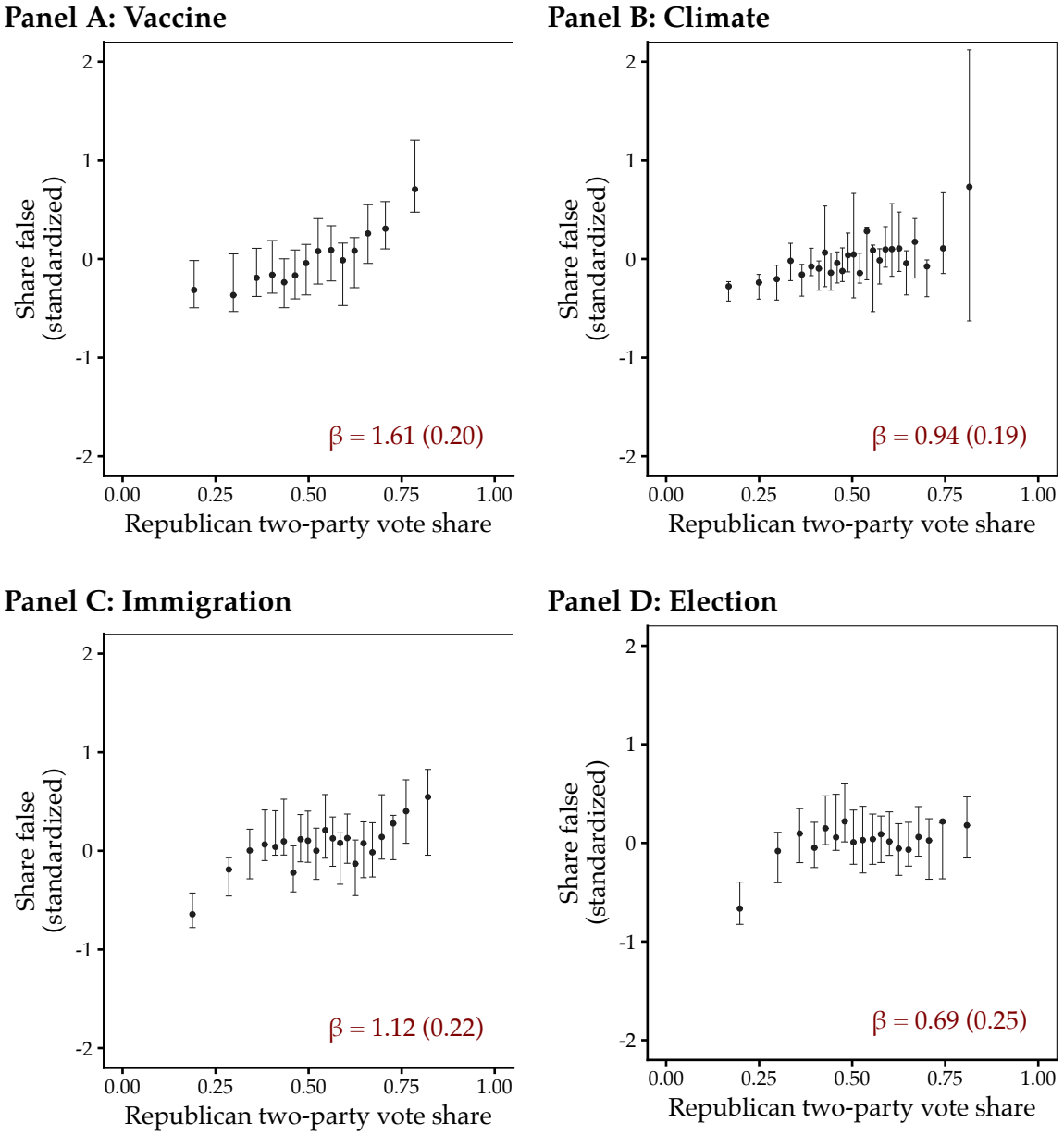
Across topics, the binscatters show that counties with higher Republican two-party vote shares tend to have higher misinformation shares, though the strength of the relationship varies by topic area. Most notably, a 1 unit increase in the Republican two-party vote share (i.e. switching from 0 percent Republican to 100 percent Republican) is associated with a 1.61 standard deviation increase in the share of vaccine-related tweets that are false, and a 1.12 standard deviation increase in the share of immigration-related tweets that are false. The relationship is weaker and noisier for climate change and election-related tweets.

Importantly, these patterns are purely descriptive: Partisanship is correlated with many other county characteristics (demographics, media environments, and patterns of Twitter use), so the figure should be interpreted as documenting a stylized fact about the cross-sectional distribution of misinformation, not as evidence of a causal effect of partisanship. At the same time, these correlations motivate the heterogeneity analyses later in the paper that examine whether the causal effects of newspaper coverage changes differ systematically across politically different places.

In addition to the relationship between political partisanship and misinformation, I also show that misinformation in one topic area tends to be correlated with misinformation in another. Figure 8 plots the six pairwise relationships between each of the four topic areas. Each observation is a county. For each county, I aggregate all tweets in a given topic and compute the share false measure, which I then standardize within each topic and exclude counties with fewer than 100 topic-specific tweets.

Notice that in panels A-C, which all have the share of vaccine-related tweets

Figure 7: Relationship between political partisanship and misinformation



Notes: This figure plots the relationship between political partisanship and misinformation in each of the four topic areas I have identified as prone to misinformation. Each observation is a county–election cycle. An election cycle is defined as the four-year period beginning in December of a presidential election year and ending in November of the subsequent presidential election year. All tweets posted within that window are aggregated to the corresponding county–election cycle. In each panel, the y-axis measures the share of tweets in a given topic area that the relevant supervised classifier predicts contains misinformation. This variable is then standardized across county-election cycles by subtracting the mean and dividing by the standard deviation. The x-axis measures the Republican two-party vote share in the most recent presidential election cycle. The univariate regression coefficient is reported in red. For each panel, I exclude from the sample all county-election cycles with fewer than 100 tweets in the relevant topic area. Binscatters are estimated using the approach outlined in Cattaneo et al. (2024). The support of the independent variable is partitioned into quantile-spaced bins and, within each bin, the conditional mean of the dependent variable is computed. The number of bins is selected in a data-dependent manner. Error bars represent pointwise 95 percent confidence intervals. Standard errors are clustered at the county level.

that are false as the independent variable, the relationships are quite weak (β between 0.04 and 0.07), suggesting that the counties that have a lot of vaccine-related misinformation are not necessarily the same as the counties that have a lot of misinformation about the other topic areas. However, panels D-F indicate that the other pairwise relationships — between the climate, immigration, and elections topic areas — are significantly stronger. For example, a 1 standard deviation increase in the share of climate-related tweets that are false is associated with a 0.31 and 0.49 standard deviation increase in the share of immigration-related and election-related tweets that are false, respectively.

5 Methodology

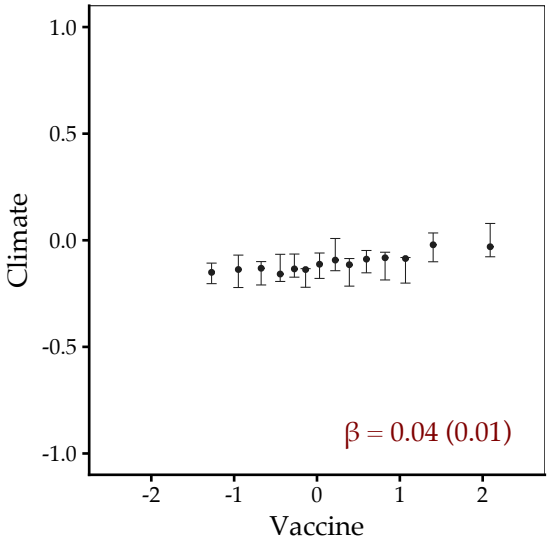
I use an event study design to determine the causal effects of local newspaper coverage changes on the volume and composition of Twitter discourse. This section discusses in detail the process used to identify and estimate these effects. Broadly, I follow the methodology outlined by Callaway and Sant'Anna (2021) for difference-in-differences with multiple periods and Callaway et al. (2024) for difference-in-differences with varying treatment intensities. I proceed as follows. First, I introduce notation. Second, I define causal parameters of interest and discuss their economic interpretation in the context of my panel. Third, I state the assumptions necessary to identify these parameters and defend their validity in the context of my panel. Fourth, I discuss potential threats to these assumptions. Finally, I explain the procedures used to estimate these parameters.

5.1 Notation

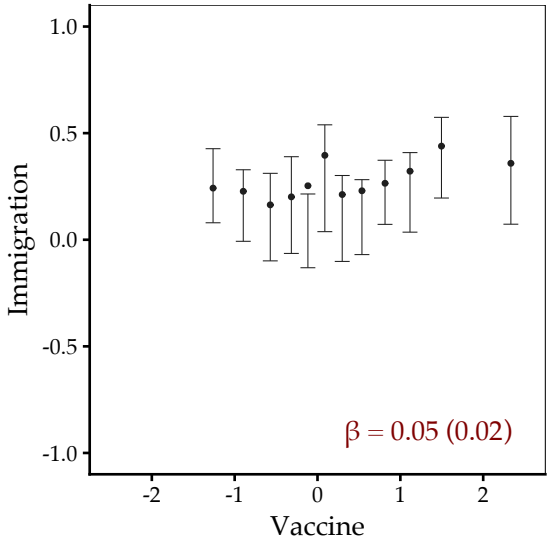
I index each county with $i \in \{1, \ldots, I\}$ and each calendar month during the sample period with $t \in \{1, \ldots, T\}$. I denote the coverage index associated with county i in period t as $Z_{i,t}$. Then a "treatment" occurs any time $Z_{i,t-1} - Z_{i,t} \neq 0$. The set of treatments associated with county i is given by $\mathcal{K}_i = \{1, \ldots, K_i\}$. Define the "dosage" of treatment k for county i as $D_{i,k} = Z_{i,t_k-1} - Z_{i,t_k}$, and let $\mathcal{D} = \{d_1, \ldots, d_J\}$ denote the support of $D_{i,k}$ across all counties and treatments, where $d_1 < \cdots < d_J$. Notice the sign on the dosage; $D_{i,k} > 0$ when the event is a *negative* coverage change (i.e.

Figure 8: Topic-level pairwise relationships in the geographic distribution of misinformation

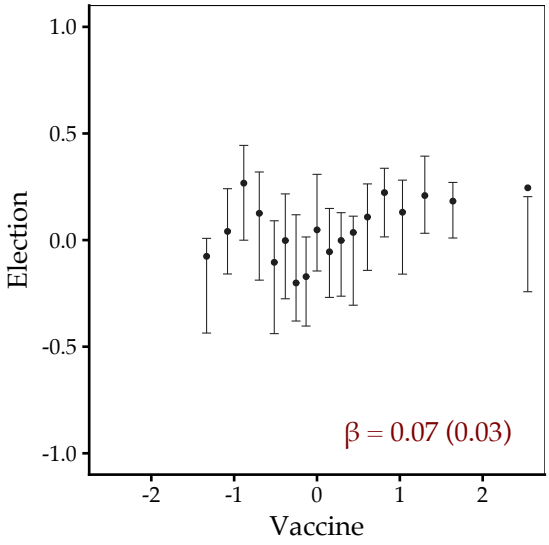
Panel A: Vaccine vs climate



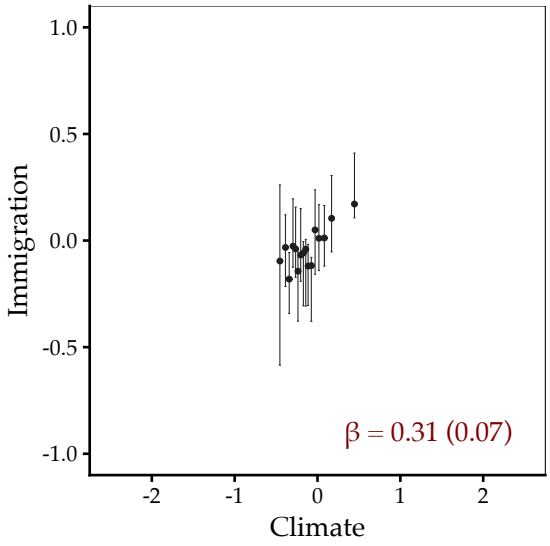
Panel B: Vaccine vs immigration

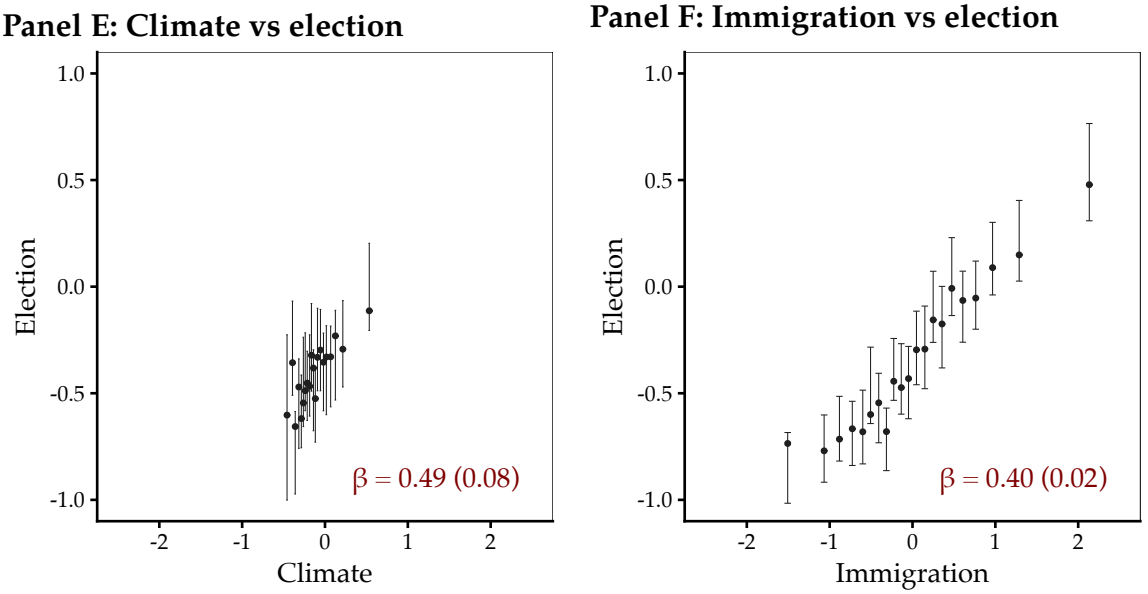


Panel C: Vaccine vs election



Panel D: Climate vs immigration





Notes: This figure plots the relationship in the geographic distribution of misinformation for each pair of topic areas. Each observation is a county. In each panel, both the x-axis and y-axis measure the share of tweets in a given topic area that the relevant supervised classifier predicts contains misinformation. This variable is then standardized across counties by subtracting the mean and dividing by the standard deviation. The univariate regression coefficient is reported in red. For each panel, I exclude from the sample all counties with fewer than 100 tweets in either of the two relevant topic areas. Binscatters are estimated using the approach outlined in Cattaneo et al. (2024). The support of the independent variable is partitioned into quantile-spaced bins and, within each bin, the conditional mean of the dependent variable is computed. The number of bins is selected in a data-dependent manner. Error bars represent pointwise 95 percent confidence intervals. Standard errors are clustered at the county level.

a county decreases newspaper coverage), and $D_{i,k} < 0$ when the event is a *positive* coverage change (i.e. a county increases newspaper coverage). Because decreases in newspaper coverage are much more common than increases over the sample period, this construction of the dosage variable makes interpreting the event study results more intuitive. Note also that $D_{i,k}$ is discrete and time-invariant.

Each county i is also associated with a set of groups $\mathcal{G}_i = \{G_{i,1}, \ldots, G_{i,K}\}$, where each $G_{i,k} \in \{1, \ldots, T\} \cup \{\infty\}$. Each group $G_{i,k}$ refers to the period in which county i is treated for the kth time. Specifically, $G_{i,k}$ is the first period in which county i experiences the updated coverage index after the kth change in coverage index. Let $G_{i,1} = \infty$ for never-treated counties.

In addition, each county i is associated with a vector of time-invariant controls $\vec{X}_i \in \mathbb{R}^p$. Importantly, the notation below requires all outcome variables and parameters of interest to be defined conditional on $\vec{X}_i$. However, to make the notation simpler, I assume in the setup that $\vec{X}_i = x$ is fixed. The necessary assumptions and implications for my empirical strategy do not meaningfully change when covariates are constant.[11]

I adopt canonical notation to describe potential outcomes. Let $Y_{i,t}(g, d)$ be the potential outcome that county i would experience in period t from being treated in period g with dosage d. However, a key complication arises when a county experiences multiple treatments over the sample period. In this case, the potential outcomes notation $Y_{i,t}(g, d)$ is under-specified, because the realized outcome in period t may depend on the county's entire treatment history, rather than any single (g, d) pair. Without strong assumptions on how multiple treatments interact, there is no well-defined $Y_{i,t}(g, d)$ for a county that is treated multiple times.

I resolve this issue by restricting attention to treatment-county pairs in which only a single treatment falls within the estimation window. Fix a window of length $2L$ months centered on the treatment date, extending L months on either side.

For each county i and treatment k, define the event window $\mathcal{W}_{i,k} = [G_{i,k} -$

[11] See section 2.3 of Callaway and Sant'Anna (2021) for a discussion of the parameters of interest and identifying assumptions when $\vec{X}_i$ is a random vector. The primary meaningful difference is the inclusion of an overlap assumption.

$L, G_{i,k} + L]$. Then define the set of *admissible* treatments for county i as

$$\mathcal{K}_i^* = \left\{k \in \mathcal{K}_i : \nexists\, k' \neq k \text{ such that } G_{i,k'} \in \mathcal{W}_{i,k}\right\}, \tag{1}$$

so that treatment k is admissible only if no other treatment — whether earlier or later — falls within its window. For each admissible pair (i, k) with $k \in \mathcal{K}_i^*$, the county's relevant treatment history within $\mathcal{W}_{i,k}$ is fully summarized by $(G_{i,k}, D_{i,k})$, and I write $Y_{i,t}(G_{i,k}, D_{i,k})$ for $t \in \mathcal{W}_{i,k}$ as the relevant potential outcome. County-treatment pairs for which $k \notin \mathcal{K}_i^*$ are excluded from the analysis entirely. This restriction ensures that the pre- and post-periods of every event window I use are free from contamination by other treatments, so that $Y_{i,t}(G_{i,k}, D_{i,k})$ is well-defined.

The observed data can therefore be represented as:

$$\{Y_{i,1}, \ldots, Y_{i,T}, G_{i,1}, \ldots, G_{i,K_i}, D_{i,1}, \ldots, D_{i,K_i}\}_{i=1}^{n} \tag{2}$$

Throughout the remainder of the paper, I omit indices i and k for ease of notation.

5.2 Definitions of causal parameters

Two causal parameters are of particular interest and policy relevance. The first is the "average treatment effect on the treated" (ATT), which is given by:

$$ATT(g, t, d) = \mathbb{E}[Y_t(g, d) - Y_t(\infty, 0) | G = g, D = d] \tag{3}$$

This parameter measures the average difference in outcomes between receiving dosage $D_i = d$ and remaining untreated, for timing group G, in period t, among events in group G that do, in fact, receive dosage d. In the context of my panel, this parameter reflects the average difference in outcomes from a county's coverage index changing by d newspaper-days, relative to a counterfactual in which coverage had remained unchanged.

The second parameter of interest is the "average causal response on the treated" $(ACRT)$, which is given by:

$$ACRT(g, t, d_j, d_{j-1}) = \mathbb{E}[Y_t(g, d_j) - Y_t(g, d_{j-1}) | G = g, D = d_j] \tag{4}$$

This parameter measures the average difference in outcomes due to a "marginal" change in dosage from d_{j-1} to d_j, for timing group g, in period t, among events in group g that do, in fact, receive dosage d_j. In the context of my panel, this parameter reflects the average difference in outcomes from a county's coverage index changing by d_j newspaper-days, relative to a counterfactual in which coverage had changed by d_{j-1} newspaper-days.

Because both $ATT(g,t,d)$ and $ACRT(g,t,d_j,d_{j-1})$ are high-dimensional, I aggregate these parameters in four ways.

(1) **Event study aggregation.** In an event study aggregation, I provide the average ATT in different periods relative to treatment. Formally, let $e = t - g$ denote the number of periods that have elapsed since unit i was treated. Then the event study aggregates for the ATT parameter are given by:

$$ATT^{es}(e) = \sum_{d \in \{\mathcal{D}: d \neq 0\}} \sum_{g \in \mathcal{G}} P(G = g, D = d | D \neq 0, G + e \leq T) \cdot ATT(g, g + e, d) \quad (5)$$

In the context of my panel, $ATT^{es}(e)$ measures the average difference in outcomes e months after a change in coverage index, relative to a counterfactual in which there was no change in coverage index, pooling across all timing groups and dosage levels. When $e \geq 0$, $ATT^{es}(e)$ can be interpreted as treatment effect dynamics. When $e < 0$, $ATT^{es}(e)$ can be used as pre-tests for the parallel trends assumption. Because I average over all dosage levels, $ATT^{es}(e)$ is equivalent to the dynamic average treatment effects on the treated if all treatments were instead defined as indicator variables, where $D_i = 1$ if county i is ever-treated, and where $D_i = 0$ if county i is never-treated.

(2) **Normalized event study aggregation.** In a normalized event study aggregation, I provide the *per-dosage* average ATT in different periods relative to treatment. This is identical to the aggregation above, except I first divide $ATT(g, g+e, d)$ by the dosage prior to aggregation:

$$ATT^{es,norm}(e) = \sum_{d \in \{\mathcal{D}: d \neq 0\}} \sum_{g \in \mathcal{G}} P(G = g, D = d | D \neq 0, G + e \leq T) \cdot \frac{1}{d} \cdot ATT(g, g + e, d) \quad (6)$$

(3) **Dosage response aggregation.** In a dosage response aggregation, I average treatment effects across all post-treatment time periods and provide the average ATT at different dosage values. Formally, we have

$$ATT^{dose}(d) = \sum_{g \in \mathcal{G}} \sum_{e=0}^{L} \frac{1}{L+1} \cdot P(G = g | D = d, G \leq T) \cdot ATT(g, g+e, d) \tag{7}$$

The parameter $ATT^{dose}(d)$ reflects the overall average treatment effect of dose d among units that are ever treated with dose d. In the context of my panel, this aggregation provides time-invariant total treatment effects for each dosage change in coverage index.

(4) **Simple aggregation.** In a simple aggregation, I average treatment effects across all post-treatment time periods and dosages into a single scalar estimate. For the ATT, this is given by

$$ATT^{simple} = \sum_{d \in \{\mathcal{D}: d \neq 0\}} \sum_{g \in \mathcal{G}} \sum_{e=0}^{L} \frac{1}{L+1} \cdot P(G = g, D = d | D \neq 0, G \leq T) \cdot ATT(g, g+e, d) \tag{8}$$

and for the $ACRT$, by

$$ACRT^{simple} = \sum_{j=1}^{J} \sum_{g \in \mathcal{G}} \sum_{e=0}^{L} \frac{1}{L+1} \cdot P(G = g, D = d_j | D \neq 0, G \leq T) \cdot \frac{ACRT(g, g+e, d_j, d_{j-1})}{d_j - d_{j-1}} \tag{9}$$

Intuitively, ATT^{simple} summarizes the overall average treatment effect, whereas $ACRT^{simple}$ summarizes the average marginal effect of an additional unit of dosage, and therefore represents the average rate of change of $ATT^{dose}(d)$ with respect to d. In the context of my panel, $ACRT^{simple}$ has an important interpretation: It is the average marginal effect — across all post-treatment periods and (d_j, d_{j-1}) pairs — of a county losing a single newspaper-day of coverage per week.

5.3 Identifying assumptions

Four assumptions are necessary to identify the ATT:

(1) **The stable unit treatment value assumption.** The potential outcomes of county i are uniquely determined by observing $D_{i,k}$ and $G_{i,k}$. There can be no spillovers from county j in group $G_{j,l}$ receiving treatment intensity $D_{j,l}$.

(2) **The random sampling assumption.** The observed data consist of

$$\{Y_{i,1}, \ldots, Y_{i,T}, G_{i,1}, \ldots, G_{i,K_i}, D_{i,1}, \ldots, D_{i,K_i}\}_{i=1}^{I} \tag{10}$$

which are independent and identically distributed.

(3) **The non-anticipation assumption.** For all $g \in \{\mathcal{G}_i\}_{i=1}^{I}$, for all $d \in \mathcal{D}$, and for all $t \in \{1, \ldots, T\}$ with $t < g$ (i.e. in pre-treatment periods), we have $\mathbb{E}[Y_{i,t}(g,d)|G = g, D = d] = \mathbb{E}[Y_{i,t}(\infty, 0)|G = g, D = d]$.

(4) **The parallel trends assumption.** For all $g \in \{\mathcal{G}_i\}_{i=1}^{I}$, for all $d \in \mathcal{D}$, and for all $t \in \{2, \ldots, T\}$, trends in never-treated potential outcomes are parallel. Formally, we have:

$$\mathbb{E}[Y_t(\infty, 0) - Y_{t-1}(\infty, 0)|G = g, D = d] = \mathbb{E}[Y_t(\infty, 0) - Y_{t-1}(\infty, 0)|G = \infty, D = 0] \tag{11}$$

An additional fifth assumption is required to identify the $ACRT$:

(5) **The strong parallel trends assumption.** For all $g \in \{\mathcal{G}_i\}_{i=1}^{I}$, for all $t \in \{2, \ldots, T\}$, and for all adjacent pairs $(d_j, d_{j-1}) \in \mathcal{D}$, trends in potential outcomes are the same regardless of the dosage actually received, conditional on being in timing group g. In other words, within events in group G, counties that received dosage d_j would experience the same outcomes as those that received dosage d_{j-1}, had such counties been assigned dosage d_{j-1}. Formally, we have:

$$\mathbb{E}[Y_t(g, d_j) - Y_{t-1}(g, d_j)|G = g, D = d_j] = \mathbb{E}[Y_t(g, d_j) - Y_{t-1}(g, d_j)|G = g, D = d_{j-1}] \tag{12}$$

5.4 Threats to identification

5.4.1 Multiple treatments

Several counties are treated multiple times. This threatens the SUTVA assumption, because the potential outcomes of county i may depend on both $(D_{i,k}, G_{i,k})$ as well as $(D_{i,k'}, G_{i,k'})$. I address this concern with the admissibility restriction $\mathcal{K}_i^*$, which ensures that no other treatment falls within the window $\mathcal{W}_{i,k}$. Consequently, the treatment history within the relevant window is fully summarized by $(G_{i,k}, D_{i,k})$.

Based on the length of my sample period, I choose a window length of 24 months on either side of the treatment date. A longer window demands more of the data, because as the window length increases, I observe data for fewer counties. However, 24 months on either side of the treatment date is a reasonable window to capture the full dynamics of treatment effects.

5.4.2 Transformations of the outcome variable

Both the parallel trends assumption and the strong parallel trends assumption are not invariant to monotonic transformations of the outcome variable. If the parallel trends assumption holds when the outcome variable is expressed in levels, it will generally not hold when the outcome variable is expressed in logs, and vice versa. In many of the analyses presented in section 6, I am interested in "count"-like outcome variables. For example, $Y_{i,t}$ could represent the number of tweets about vaccines posted by users in county i in period t.

For such outcomes, I argue that parallel trends in logs is more appropriate than parallel trends in levels. Parallel trends is an assumption about untreated potential outcomes $Y_t(\infty, 0)$, and requires that, absent treatment, all counties trend the same way over time. The relevant question is therefore how common shocks propagate through the outcome variable. Counties vary enormously in population, urbanism, Twitter penetration, and baseline engagement with social media. As a result, I argue that such shocks — say, a global pandemic that elevates the salience of vaccine-related discourse — are more likely to scale engagement *proportionally* with baseline activity rather than adding a *fixed number of tweets* uniformly across counties. Under this view, common shocks like the Covid-19 pandemic operate multiplicatively,

which means that parallel trends is more naturally satisfied on a log scale than in levels. If I instead assumed that parallel trends held in levels, I would require that these common shocks shift tweet volumes by *the same absolute number* regardless of whether a county has 100 or $100,000$ baseline monthly tweets — a considerably less plausible restriction.

However, specifying parallel trends in logs introduces a practical complication: The natural logarithm function is undefined at zero. Many of the "count"-like outcome variables used in the analyses in section 6 are often equal to zero for certain months in certain counties. To resolve this issue, I adopt an approach described in Chen and Roth (2024). Rather than applying the natural log function directly, I apply the following "modified log" function:

$$m(Y_{i,t}) = \begin{cases} \log(Y_{i,t}) & ; \quad Y_{i,t} > 0 \\ \log\left(\min_{Y>0}\{Y\}\right) & ; \quad Y_{i,t} = 0 \end{cases} \tag{13}$$

The "modified log" function behaves identically to the natural log for the majority of its domain. However, when the outcome variable is zero, the modified log returns the log of the minimum non-zero value of the outcome variable. Intuitively, in applying this transformation, I pretend that counties with a zero-valued outcome variable instead have the minimum non-zero value of the outcome variable. This function is a principled approach to handling zero-valued outcomes, because it ensures that estimates remain interpretable. To understand why, consider that treatment could affect tweet volumes through two distinct channels: an intensive margin (changing potential outcomes for county-months that would have non-zero potential outcomes regardless of treatment) and an extensive margin (changing potential outcomes for county-months that would have outcomes of zero if it were not treated). Any transformation that assigns a finite value to zeros implicitly encodes an assumption about how to weight changes along the extensive margin relative to changes along the intensive margin.

Because my transformation sets all zeros equal to the natural logarithm of the smallest positive value observed in the data, I effectively "shut off" my ability to detect any treatment effects along the extensive margin. For example, suppose in the untransformed data, a county moved from an outcome of zero in period $t -$

1 to $\min_{Y>0}\{Y\}$ tweets in period t. In the untransformed data, there is an effect along the extensive margin. However, my transformation functionally pretends that such a county has $\min_{Y>0}\{Y\}$ outcomes all along. The resulting estimates thus mechanically set the effect along the extensive margin equal to zero and isolate only the intensive margin. They measure how changes in newspaper coverage affect outcomes only among county-months that already have non-zero outcomes.

The choice to focus entirely on the intensive margin is appropriate in my setting for two reasons. For most values of the "count"-like outcome variables in my panel, the intensive margin is the policy-relevant channel. For example, suppose $Y_{i,t}$ measures the number of tweets about vaccines posted in county i in period t. The concern motivating this paper is not that changes in newspaper coverage induce residents of county i to tweet about vaccines when they previously did not. Rather, it is that changes in newspaper coverage amplify salient topics that are prone to misinformation in communities where such content *already circulates*. I am interested in the change from $1,000$ monthly tweets about vaccines to $10,000$ monthly tweets about vaccines, not the change from zero tweets about vaccines to one tweet about vaccines. Estimates that isolate the intensive margin speak directly to this concern.

5.4.3 Twitter policy changes

The final threat to identification comes from the fact that the variables I use in my analyses presented in section 6 are constructed from Twitter activity. As demonstrated in figure 1, Twitter activity responds dramatically to platform policy changes that are entirely unrelated to local newspaper coverage. The parallel trends assumption is violated if the effects of these policy changes on tweet volumes are correlated with treatment timing or dosage. For instance, switching geotagging from opt-out to opt-in may disproportionately affect users that are also more likely to be located in counties that experience changes in local newspaper coverage. In that case, treated and control counties would follow divergent trends in untreated potential outcomes $Y_t(\infty, 0)$, and the parallel trends assumption would fail.

I address this issue by partitioning the sample period into the three distinct regimes defined in section 3.1. Formally, let $\mathcal{R}(t) \in \{R_1, R_2, R_3\}$ denote the regime associated with calendar period t. Notice that all four types of aggregations — event

study, normalized event study, dosage response, and simple — are computed by averaging over individual $ATT(g,t,d)$ cells. Each individual cell is a difference-in-differences comparison. I take the average change in outcomes for counties treated at time g with dosage d, and subtract the average change in outcomes of never-treated counties. When forming these averages, I only include (g,t) cells where g and t belong to the same regime (i.e. $\mathcal{R}(g) = \mathcal{R}(t)$). Any cell where g and t straddle a regime boundary — meaning a platform policy change occurred somewhere between the treatment date and the calendar period being compared — is not included in the aggregation. This procedure ensures that Twitter's policy changes cannot generate spurious trends in the estimated parameters, because ever-treated and never-treated control counties are never compared across a period in which such a policy change occurred.

This approach comes with its own set of drawbacks. For one, it assumes that any violations in the parallel trends assumption occur during the period in which the policy change is introduced. It does not account for the possibility that a platform policy change has persistent differential effects that propagate forward within a regime. If the effects of Twitter's policy changes on tweet volumes are correlated with the treatment, and if these effects persist throughout the regime, then removing across-regime comparisons does not remedy the problem. The within-regime comparisons remain contaminated.

A second, more mechanical drawback is that restricting to within-regime comparisons reduces the number of (g,t) cells that enter each aggregation. Fewer comparisons means less variation is used in estimation, and standard errors on the aggregated parameters are correspondingly larger.

5.5 Estimation

I estimate the causal parameters defined above in two stages. In the first stage, I estimate the disaggregated $ATT(g,t,d)$ cells. In the second stage, I aggregate these cells into the event study, dosage response, and simple estimates described above.

First, because dosage is discrete in the context of my panel, I estimate $ATT(g,t,d)$ separately for each dosage level $d \in \mathcal{D}$. Within each dosage stratum, the estimation procedure reduces to the standard staggered adoption difference-in-differences de-

sign described by Callaway and Sant'Anna (2021). I compare outcomes for counties with $G = g$ and $D = d$ to outcomes for never-treated counties with $G = \infty$ and $D = 0$. From the parallel trends assumption, we have:

$$ATT(g, t, d) = \mathbb{E}[Y_t - Y_{t-h} \mid G = g, D = d] - \mathbb{E}[Y_t - Y_{t-h} \mid G = \infty, D = 0] \quad (14)$$

The first term is the average change in outcomes between periods $t - h$ and t for counties first treated in period g with dosage d. The second term is the same average change over the same interval for never-treated counties.

I use the doubly-robust estimator proposed in Callaway and Sant'Anna (2021) to estimate $ATT(g, t, d)$. For each (g, t, d) cell, this estimator takes the form:

$$\widehat{ATT}(g, t, d) = \frac{1}{I} \sum_{i=1}^{I} \left[\left(\frac{\mathbb{1}\{G_i = g, D_i = d\}}{\hat{P}(G_i = g, D_i = d)} - \frac{\hat{p}_{g,d}(X_i)}{1 - \hat{p}_{g,d}(X_i)} \cdot \frac{\mathbb{1}\{D = 0\}}{\hat{P}(D = 0)} \right) (Y_{i,t} - Y_{i,g-1} - \hat{m}_{g,d}(X_i)) \right] \quad (15)$$

where i indexes counties in the sample, $\hat{P}(\cdot)$ denotes sample frequency estimates of group probabilities, $\hat{p}_{g,d}(X_i)$ is an estimated propensity score for being in dosage-timing group (g, d), and $\hat{m}_{g,d}(X_i)$ is an estimated outcome regression for the comparison group's change in outcomes. The doubly-robust estimator is consistent if either $\hat{p}_{g,d}$ or $\hat{m}_{g,d}$ is correctly specified, but not necessarily both. I describe the specific vector of covariates X_i in section 6. Repeating this procedure across all admissible (g, t, d) triples yields a full set of disaggregated first-stage estimates. These cells are the building blocks for all aggregated parameters.

To estimate $ATT^{es}(e)$, I average the $\widehat{ATT}(g, t, d)$ estimates across all groups and dosages, weighting each estimate based on the sample frequencies of each (g, d) group observed at event period e. Formally, we have:

$$\widehat{ATT}^{es}(e) = \sum_{d \in \{\mathcal{D}: d \neq 0\}} \sum_{g \in \mathcal{G}} \hat{P}(G = g, D = d \mid D \neq 0, G + e \leq T) \cdot \widehat{ATT}(g, g + e, d) \quad (16)$$

The normalized event study parameter is estimated analogously.

The dosage response parameter $ATT^{dose}(d)$ is estimated as follows. Within each

dosage stratum, I average the first-stage estimates across timing groups and post-treatment event periods, weighting each timing group equally. Formally:

$$\widehat{ATT}^{dose}(d) = \sum_{g \in \mathcal{G}} \sum_{e=0}^{L} \frac{1}{L+1} \cdot \hat{P}(G = g \mid D = d, G \leq T) \cdot \widehat{ATT}(g, g+e, d) \tag{17}$$

The simple aggregation ATT^{simple} then averages $\widehat{ATT}(g,t,d)$ across all groups, dosages, and event periods, weighting by the sample frequency of each dosage among ever-treated counties:

$$\widehat{ATT}^{simple} = \sum_{d \in \{\mathcal{D}: d \neq 0\}} \sum_{g \in \mathcal{G}} \sum_{e=0}^{L} \frac{1}{L+1} \cdot \hat{P}(G = g, D = d | D \neq 0, G \leq T) \cdot \widehat{ATT}(g, g+e, d) \tag{18}$$

Finally, because dosage is discrete, and assuming the strong parallel trends assumption holds, the $ACRT(g,t,d_j,d_{j-1})$ parameters can be directly computed as differences between estimates for $ATT(g,t,d_j)$ and $ATT(g,t,d_{j-1})$:

$$\begin{aligned}
ACRT(g,t,d_j,d_{j-1}) &= \mathbb{E}[Y_t(g,d_j) - Y_t(g,d_{j-1}) \mid G = g, D = d_j] && (19)\\
&= \mathbb{E}[Y_t(g,d_j) - Y_t(\infty,0) \mid G = g, D = d_j] \\
&\quad - \mathbb{E}[Y_t(g,d_{j-1}) - Y_t(\infty,0) \mid G = g, D = d_j] && (20)\\
&= \mathbb{E}[Y_t(g,d_j) - Y_t(\infty,0) \mid G = g, D = d_j] \\
&\quad - \mathbb{E}[Y_t(g,d_{j-1}) - Y_t(\infty,0) \mid G = g, D = d_{j-1}] && (21)\\
&= ATT(g,t,d_j) - ATT(g,t,d_{j-1}) && (22)
\end{aligned}$$

where the first equality is the definition of $ACRT(g,t,d_j,d_{j-1})$, and the third equality is implied by the strong parallel trends assumption.

I therefore estimate $ACRT(g,t,d_j,d_{j-1})$ by plugging in the corresponding first-stage estimates:

$$\widehat{ACRT}(g,t,d_j,d_{j-1}) = \widehat{ATT}(g,t,d_j) - \widehat{ATT}(g,t,d_{j-1}) \tag{23}$$

The aggregated parameter $\widehat{ACRT}^{simple}$ is then computed by applying the same

weighting schemes described above to these disaggregated cells.

I conduct inference using a non-parametric bootstrap procedure. For each bootstrap draw, I resample counties with replacement, re-estimate the full sequence of first-stage $\widehat{ATT}(g,t,d)$ cells, and recompute each aggregated parameter. I repeat this process for $\mathcal{B} = 1,000$ draws. Notice that standard errors for all estimates are clustered at the county level, because I resample counties (rather than county-month observations) in each bootstrap draw.

5.6 Simulation exercises

A natural concern is whether the above methodology introduces artifacts that could be mistaken for genuine treatment effects. I address this concern by conducting a simulation exercise that computes event study aggregations of ATTs, but with simulated data replacing the observed data for the outcome variable. In the simulation, the true treatment effect is known to be exactly zero. If the above methodology is well-behaved, I should recover null effects.

I first select two relevant outcome variables — (i) the modified log of the number of tweets and (ii) the share of tweets about vaccines that contain misinformation. To construct the simulated data, I take the outcome variable for each ever-treated county-month during the post-treatment period, and I replace it with a realization of the dependent variable drawn from the never-treated counties during the corresponding month. In particular, the replacement outcome for ever-treated county i in month t is from the never-treated county that has the most similar pre-treatment mean outcome to county i. Aside from these changes to the outcome variables, all other variables — including treatment group assignment and treatment dosage — are held fixed exactly as observed in the data.

Because the simulated post-treatment outcomes for ever-treated counties are drawn from the never-treated county that most closely resembles them in the pre-treatment period, the true treatment effect is zero by construction. Intuitively, my simulation verifies that there is a null effect when ever-treated counties have outcomes that look almost exactly like their closest never-treated analog after treatment.

Figure 9 presents results from this exercise. In Panel A, the outcome variable is

the modified log of the total number of tweets. In Panel B, the outcome variable is the share of tweets about vaccines that contain misinformation. In both panels, the left figure uses the simulated outcomes directly. The right figure adds a constant of 1 to the simulated outcome.

Table 3 presents simple aggregates of these simulated treatment effects. In both cases, the event study aggregates of the ATT return precisely the expected effects; null in the left column, and precisely 1 in the right column. These simulation exercises validate that the estimation procedure does not mechanically generate spurious treatment effects. When the true effect is zero, the estimator correctly returns estimates near zero with well-behaved confidence intervals.

Table 3: Simple aggregates of ATTs, simulated outcomes

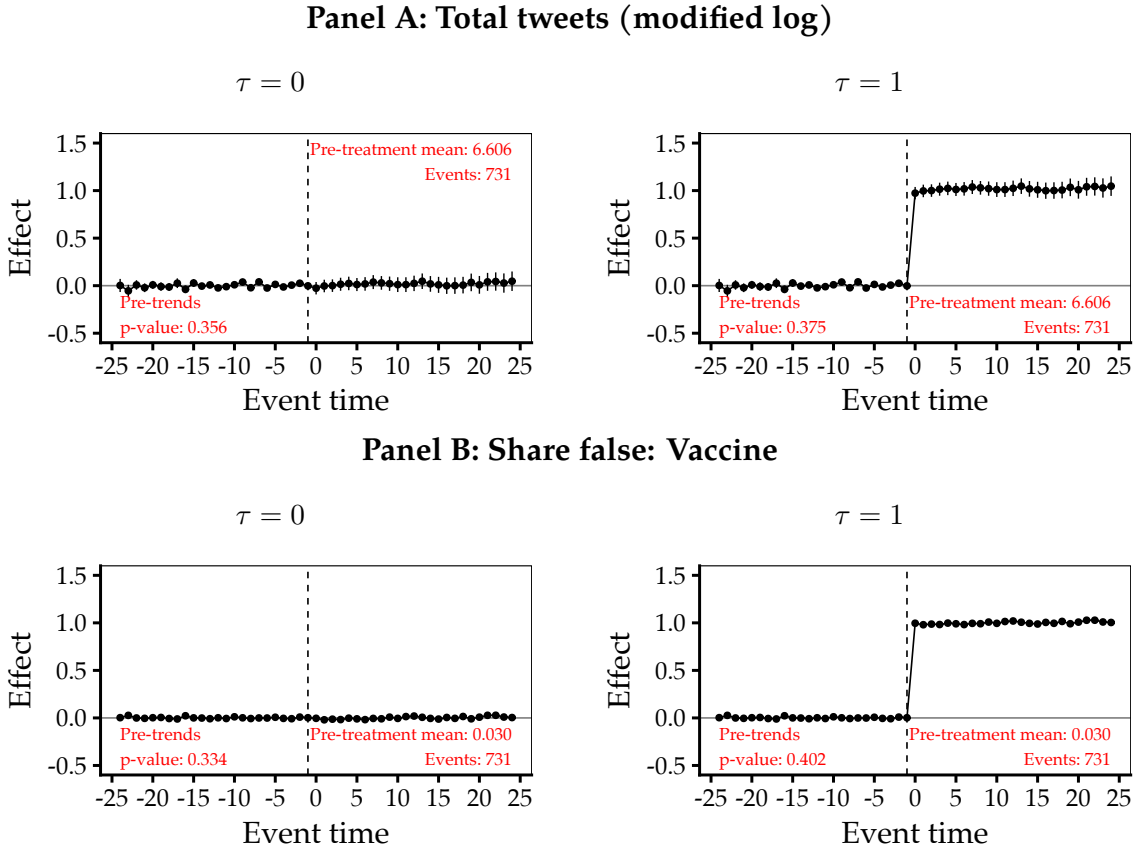
Treatment effect	$\tau = 0$	$\tau = 1$
Total tweets (modified log)	0.018	1.018
	(0.041)	(0.041)
Share false: Vaccine	0.000	1.000
	(0.006)	(0.006)

Notes: This table provides simple aggregates of ATTs for simulated outcomes over the 24 months (2 years) before and after a county-level coverage change. The ATT parameter and aggregation procedure is described in detail in section 5 of the text. The simulation procedure is described in detail in section 5.6 of the text. In the first row, the dependent variable is the modified log of the total number of tweets posted in each county in each month. In the second row, the dependent variable is the share of tweets about vaccines that the relevant supervised classifier labels as likely to contain misinformation. Corresponding event study aggregations can be found in figure 9. I exclude from the sample all ever-treated and never-treated counties with fewer than 100 tweets in the relevant topic area over the entire sample period. Standard errors are clustered at the county level.

6 Estimates of causal effects

In this section, I estimate several causal effects of exposure to local newspapers. First, I show that when a local newspaper changes its coverage frequency, there are, in fact, meaningful changes in the number of articles published each month. Second, I present coefficients from a naive regression that estimates the cross-sectional relationship between a county's average coverage index and both the volume and composition of its Twitter activity. These cross-sectional estimates serve as useful benchmarks against which to compare the results of the difference-in-differences

Figure 9: Event study aggregates of ATTs, simulated outcome variables



Notes: This figure plots "event study" aggregates of ATTs for the 24 months (2 years) before and after a county-level coverage change. The ATT parameter and aggregation procedure is described in detail in section 5 of the text. The simulation procedure is discussed in section 5.6 of the text. In panel A, the dependent variable is the modified log of the total number of tweets posted in each county in each month. In panel B, the dependent variable is the share of tweets about vaccines that the relevant supervised classifier labels as likely to contain misinformation. The left figure uses the simulated outcomes directly. The right figure adds a constant of 1. Annotations are provided in red to contextualize treatment effects for each figure. I pool the dependent variable for all ever-treated counties across all pre-treatment periods and report the mean. The number of "events" refers to the number of county-level changes in coverage that are used in estimation. The test for pre-trends is computed by conducting a Wald test using the bootstrapped covariance matrix. Error bars represent pointwise 95 percent confidence intervals. Standard errors are clustered at the county level.

specification. Third, I demonstrate that coverage changes have precise null effects on the *volume* of Twitter activity across all topic areas. Fourth, I show that coverage changes affect the *composition* of Twitter activity about vaccines and immigration. Fifth, I show that treatment effects vary based on treatment dosage (at least for several topic areas) and are increasing in dosage size for the composition of vaccine-related and immigration-related Twitter activity. Finally, I show that treatment effects on the composition of Twitter activity about vaccines and immigration are roughly twice as large for Democratic counties as they are for Republican counties.

For all event study aggregations of the treatment effects discussed above, I examine whether I observe pre-trends in the estimated treatment effects during the periods prior to treatment. Because treatment has not yet occurred, estimated effects during the pre-treatment periods should be statistically indistinguishable from zero. If they are non-zero, the parallel trends assumption is violated. To test for pre-trends, I conduct a Wald test that assesses whether all pre-treatment effects are jointly equal to zero. If I fail to reject the null, I fail to reject a key testable implication of the parallel trends assumption.

6.1 Treatment effects on article output

First, I estimate the causal effects of coverage changes on newspapers' article output. The goal of this exercise is to demonstrate a kind of "first-stage" effect of treatment. This exercise shows that coverage changes do, in fact, affect the number of articles that newspapers publish and thus meaningfully impact local news coverage.

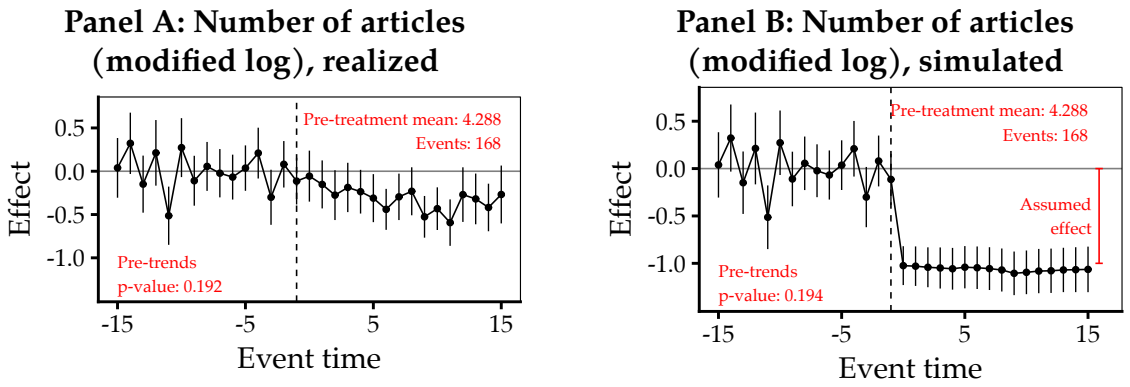
To do this, I use the number of articles published each month on NewsLibrary.com as the outcome of interest, which serves as a proxy for a newspaper's article output. Because the number of articles published each month is a "count"-like variable, I define the dependent variable to be the modified log of the number of articles published on NewsLibrary.com by a given newspaper in a given month.

Then, I deploy the methodology described in section 5. Each unit is a distinct newspaper (as opposed to a distinct county). Treated units consist of all newspapers with coverage changes at some point during the sample period that have article

output data scraped from NewsLibrary.com. Control units consist of never-treated newspapers that have article output data scraped from NewsLibrary.com.

Figure 10 presents event study aggregates of ATTs for the 15 months before and after a newspaper makes a coverage change. Panel A presents realized treatment effects. Because the dependent variable is the modified log of the number of articles, these effects can be interpreted as percent changes along the intensive margin. In other words, a coverage change induces a decrease in article output by approximately 20 percent for ever-treated newspapers over the course of the post-treatment period, conditional on the newspaper publishing *any* articles on NewsLibrary.com.

Figure 10: Event study aggregates of ATTs, number of articles published on NewsLibrary.com



Notes: This figure plots "event study" aggregates of ATTs for the 15 months before and after a newspaper-level coverage change. The ATT parameter and aggregation procedure is described in detail in section 5 of the text. Panel A provides treatment effects of a change in coverage on the number of published articles scraped from NewsLibrary.com. The dependent variable is the modified log of the total number of published articles scraped from NewsLibrary.com for a given newspaper in a given month. Panel B provides analogous treatment effects under a simulation in which the number of published articles scraped from NewsLibrary.com is mechanically set equal to zero for each period post-treatment. Annotations are provided in red to contextualize treatment effects for each figure. I pool the dependent variable for all ever-treated newspapers across all pre-treatment periods and report the mean. The number of "events" refers to the number of newspaper-level changes in coverage that are used in estimation. The test for pre-trends is computed by conducting a Wald test using the bootstrapped covariance matrix. For each panel, I exclude from the sample all ever-treated and never-treated newspapers with fewer than 100 published articles scraped from NewsLibrary.com over the entire sample period. Error bars represent pointwise 95 percent confidence intervals. Standard errors are clustered at the county level.

Panel B presents treatment effects for a hypothetical scenario in which all ever-treated newspapers eliminated article output on NewsLibrary.com entirely. In this

scenario, article output is mechanically set to zero for all ever-treated newspapers during the post-treatment period. As expected, under this hypothetical scenario, article output for ever-treated newspapers declines by 100 percent.

The discrepancy between panel A and panel B suggests that, on average, treatment reduces article output by 20 percent of a full newspaper closure. This makes sense given that the majority of treatments for newspapers with articles published on NewsLibrary.com are not complete closures. Rather, the average treatment intensity is 3.76.

6.2 Cross-sectional estimates

Next, I present cross-sectional estimates for the relationship between a county's newspaper coverage and both the volume and composition of Twitter discourse. The goal of this exercise is not to estimate causal effects, but rather to demonstrate the relationship in the cross-section and provide an additional benchmark for the causal effects estimated in sections 6.3 and 6.4.

To do this, I estimate:

$$Y_i = \alpha + \beta \text{mean\_coverage\_index}_i + \varepsilon \tag{24}$$

where i indexes ever-treated counties, and the independent variable is the average coverage index for a given county across the sample period.

Table 4 presents regression estimates for β for eight different outcome variables. In panel A, the outcome variable is the modified log of the total number of tweets about a given topic, aggregated across the sample period. These estimates measure the relationship between the coverage index and tweet volume. For all topic areas, the relationship is positive and statistically significant. Because the outcome variable is in (modified) logs, the relationship can be interpreted as percent changes along the intensive margin. A one-unit increase in coverage index is associated with a 3.7, 3.5, 4.2, and 4.2 percent increase in the number of tweets about vaccines, climate change, immigration, and U.S. election results respectively. This positive relationship makes sense, considering that there are both more newspapers and more Twitter activity in more populous places.

Table 4: Cross-sectional relationship between newspaper coverage and Twitter activity

	Vaccine	Climate	Immigration	Election
Panel A: Total tweets (modified log)				
Coverage index	0.037	0.035	0.042	0.042
	(0.007)	(0.007)	(0.009)	(0.008)
Observations	337	342	414	396
Panel B: Share false				
Coverage index	-0.023	-0.005	-0.031	-0.009
	(0.009)	(0.005)	(0.011)	(0.006)
Observations	337	342	414	396

Notes: This table provides cross-sectional estimates for the relationship between a county's newspaper coverage and its Twitter activity. Each observation is an ever-treated county. The independent variable is the coverage index of a given county averaged across the entire sample period. The dependent variable in panel A is the modified log of total tweets about a given topic over the full sample period. The dependent variable in panel B is the share of tweets about a given topic that the relevant supervised classifier labels as likely to contain misinformation. I exclude from the sample all ever-treated counties with fewer than 100 tweets in the relevant topic area.

In panel B, the outcome variable is the share of tweets about a given topic that contain misinformation. These estimates measure the relationship between the coverage index and the composition of Twitter discourse. The only coefficients significant at the 95 percent level are the ones for vaccine-related tweets and immigration-related tweets. A one unit increase in a county's average coverage index is associated with a 2.3 percentage point decrease in the share of false content about vaccines, and a 3.1 percentage point decrease in the share of false content about immigration.

6.3 Treatment effects on the volume of discourse on Twitter

Next, deploying the methodology in section 5, I estimate treatment effects on the volume of Twitter activity. I measure Twitter activity in three ways. The first — and preferred — outcome is simply the number of tweets published by users in county i in month t. The second outcome is the total number of likes that these tweets received. The third outcome is the total number of reposts that these tweets receive. Event study aggregates for total likes and total reposts are presented in

appendix C.

I prefer the first metric for several reasons. Most importantly, likes and reposts are relatively sparse in the data. According to table 2, the median tweet in the Geotweet Archive receives zero likes and zero reposts. As a result, estimates are highly sensitive to the few tweets with a large number of likes and reposts. In addition, I observe the location of the user who initially posted the tweet, but I do not observe the location of the users who liked or reposted the tweet. As a result, tweets about vaccines, say, could accumulate many likes and reposts due to the behavior of users in a completely different part of the country who are entirely unaffected by the local newspaper coverage change in county i. In contrast, the decision to post the tweet initially is made entirely by a user who would, in theory, be affected by the coverage change.

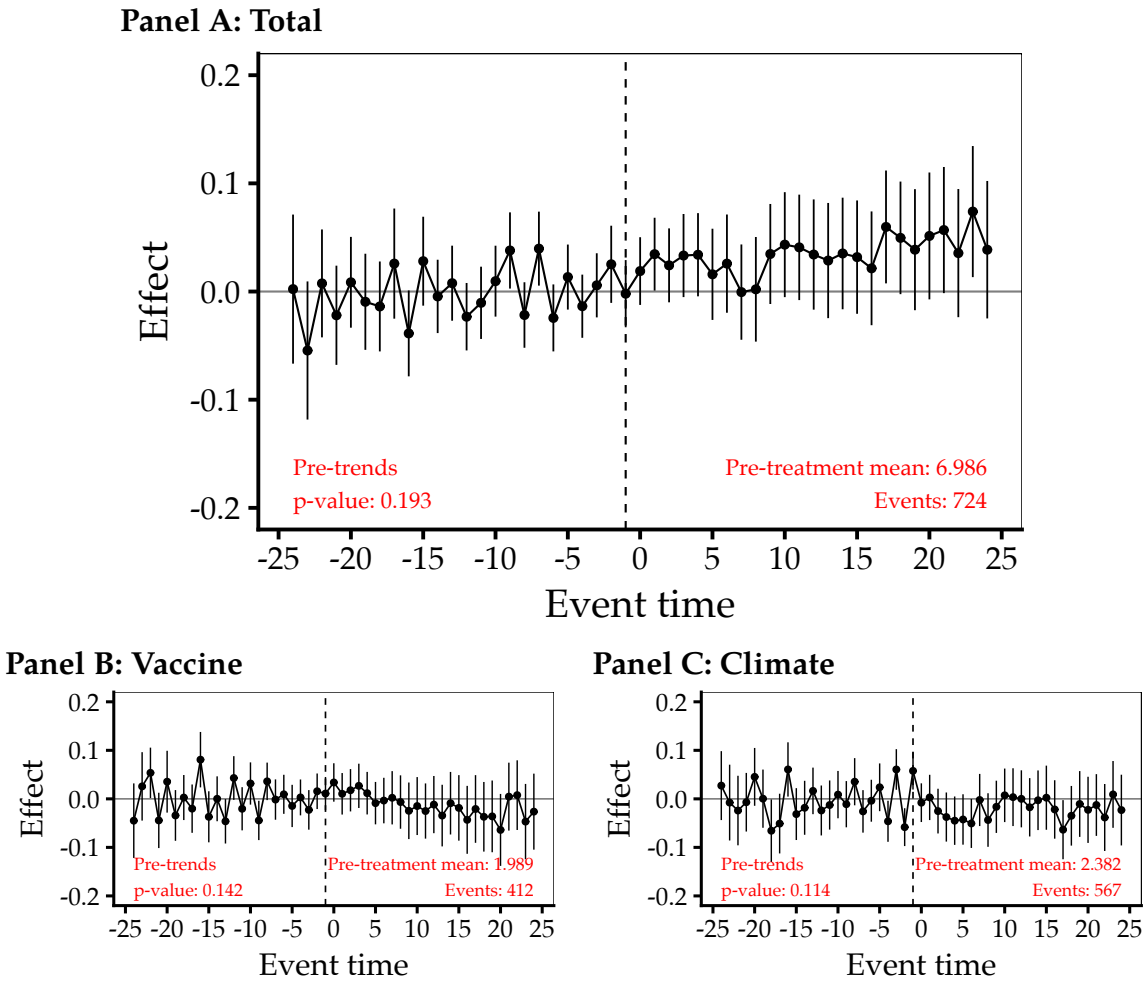
Because all of the above outcomes are "count"-like, I define $Y_{i,t}$ to be the modified log of the number of tweets for county i in period t. In addition, I include, as a time-invariant control, the population of county i in 2013, at the beginning of the sample period.

Figure 11 presents treatment effects on the volume of Twitter activity. In panel A, the outcome variable is the modified log of the total number of tweets. In panels B-E, the outcome variable is the modified log of the total number of tweets about each of the four topic areas I have identified as prone to misinformation. In panels F and G, the outcome variable is the modified log of the total number of tweets about each of the two placebo topic areas I have identified as likely to remain unaffected by changes in local newspaper coverage.

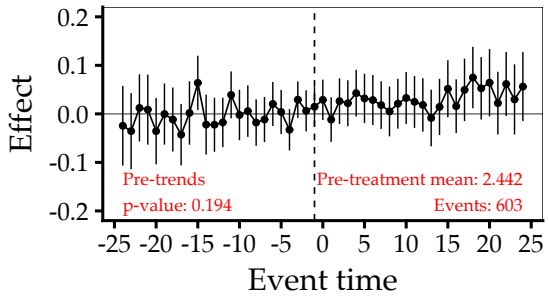
Importantly, I exclude from my sample all ever-treated and never-treated counties with fewer than 100 tweets that pertain to the relevant topic area over the entire sample period. When baseline tweet counts are very small, trivial changes in levels (which could be entirely driven by noise) can translate into large percentage changes. This restriction ensures that my analysis focuses on counties with a meaningful baseline level of Twitter activity in each topic area.

Because the dependent variable is in (modified) logs, these effects can be interpreted as percent changes along the intensive margin. All panels report precise null effects. The aggregated ATT during almost all months relative to treat-

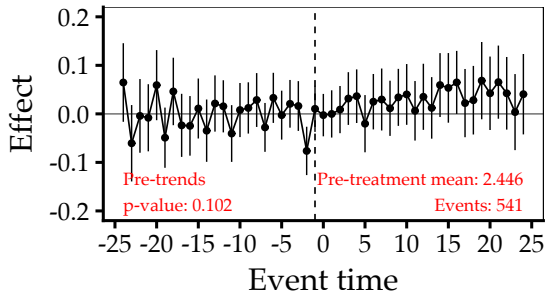
Figure 11: Event study aggregates of ATTs, total tweets



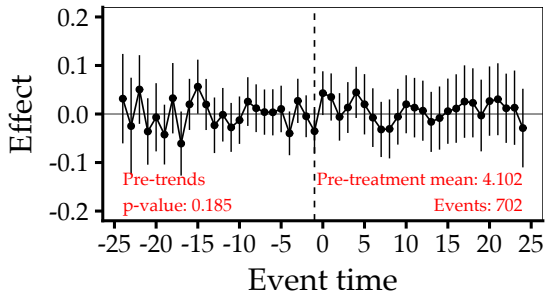
Panel D: Immigration



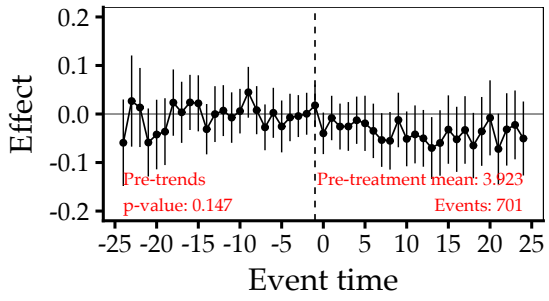
Panel E: Election



Panel F: Sports



Panel G: Music



Notes: This figure plots "event study" aggregates of ATTs for the 24 months (2 years) before and after a county-level coverage change. The ATT parameter and aggregation procedure is described in detail in section 5 of the text. Panel A provides event study plots for overall Twitter activity. The dependent variable is the modified log of the total number of tweets posted in a given county in a given month. The modified log function is described in section 5 of the text. Panels B-E provide analogous event study plots, except the dependent variables only include tweets that pertain to topic areas I have identified as prone to misinformation. Panels F and G provide analogous event study plots, except the dependent variables only include tweets that pertain to the two placebo topic areas I have identified as likely to remain unaffected by changes in newspaper coverage. Annotations are provided in red to contextualize treatment effects for each figure. I pool the dependent variable for all ever-treated counties across all pre-treatment periods and report the mean. The number of "events" refers to the number of county-level changes in coverage that are used in estimation. The test for pre-trends is computed by conducting a Wald test using the bootstrapped covariance matrix. For each panel, I exclude from the sample all ever-treated and never-treated counties with fewer than 100 tweets in the relevant topic area over the entire sample period. Error bars represent pointwise 95 percent confidence intervals. Standard errors are clustered at the county level.

ment hovers around zero, and confidence intervals generally extend only a few hundredths of a percentage point in the positive or negative direction. This implies that coverage changes cause a zero percent change (approximately) in the number of tweets posted within ever-treated counties, conditional on the county already experiencing some non-zero amount of Twitter activity. These null effects are persistent across topic areas. In addition, treatment effect dynamics do not look meaningfully different in panels B-E (topics of interest) compared to panels F and G (placebo topics). In addition, all panels fail to reject the null hypothesis of no pre-trends.

Table 5 presents simple aggregates of treatment effects. These provide pointwise estimates of effects across the entire 24-month post-treatment period. All estimates in the left column — my preferred metric — are statistically insignificant at the 95 percent level. These null estimates are in tension with the positive cross-sectional estimates presented in section 6.2, likely because the difference-in-differences specification avoids concerns about omitted variables that prevent the cross-sectional estimates from having a causal interpretation.

6.4 Treatment effects on the composition of discourse on Twitter

I now estimate treatment effects on the composition of discourse on Twitter rather than the overall volume of discourse. I define $Y_{i,t}$ to be the share of tweets posted by users in county i during period t that the relevant supervised classifier predicts to contain misinformation. As in section 6.3, I include the population of each county in 2013 as a control.

Figure 12 presents event study aggregates of ATTs for the 24 months $(2$ years$)$ before and after a county experiences a coverage change. Each panel provides estimates for one of the four topic areas I have identified as prone to misinformation. I do not report analogous panels for the two placebo topics, because the outcome in this specification relies on a meaningful distinction between "true" and "false" content. There is no comparable classification framework that allows tweets about either sports or music to be labeled as true or false in a systematic way.

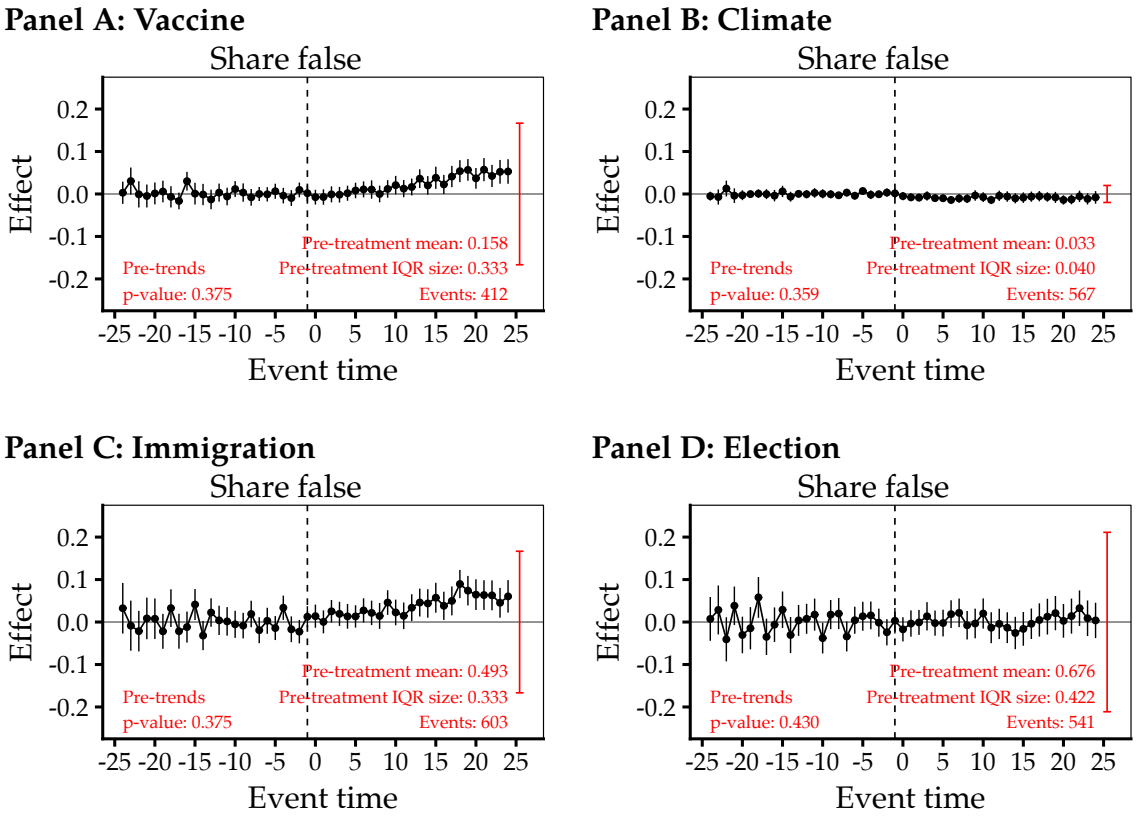
I estimate positive and significant effects of a coverage change on the share of false tweets about vaccines (panel A) and the share of false tweets about immigra-

Table 5: Simple aggregates of ATTs, tweet volumes

Treatment effect	Total tweets	Total likes	Total retweets
All tweets	0.035	0.101	0.041
	(0.028)	(0.047)	(0.047)
Vaccine	-0.013	-0.020	-0.013
	(0.026)	(0.009)	(0.006)
Climate	-0.020	-0.022	-0.017
	(0.026)	(0.014)	(0.010)
Immigration	0.031	-0.023	-0.012
	(0.027)	(0.020)	(0.017)
Election	0.030	-0.024	-0.018
	(0.030)	(0.019)	(0.016)
Sports	0.025	0.001	-0.015
	(0.032)	(0.033)	(0.034)
Music	-0.040	-0.042	-0.010
	(0.030)	(0.030)	(0.029)

Notes: This table provides simple aggregates of ATTs for the 24 months (2 years) before and after a county-level coverage change. The ATT parameter and aggregation procedure is described in detail in section 5 of the text. The dependent variable in the left column is the modified log of the number of tweets posted in a given county in a given month. The dependent variable in the middle column is the modified log of the total number of likes that these tweets received. The dependent variable in the right column is the modified log of the total number of reposts that these tweets received. Each row contains the treatment effects for the given topic area. Corresponding event study aggregations can be found in figures 11, 17, and 18. I exclude from the sample all ever-treated and never-treated counties with fewer than 100 tweets in the relevant topic area over the entire sample period. Standard errors are clustered at the county level.

Figure 12: Event study aggregates of ATTs, share of tweets containing misinformation



Notes: This figure plots "event study" aggregates of ATTs for the 24 months (2 years) before and after a county-level coverage change. The ATT parameter and aggregation procedure is described in detail in section 5 of the text. The dependent variable is the share of tweets posted in a given county in a given month that the relevant supervised classifier labels as likely to contain misinformation. Each panel provides event study estimates for one of the four topic areas I have identified as prone to misinformation. Annotations are provided in red to contextualize treatment effects for each figure. I pool the dependent variable for all ever-treated counties across all pre-treatment periods and report both the mean and difference between the 25th percentile value and the 75th percentile value. The size of the interquartile range is plotted, centered at zero, on the right side of each figure. The number of "events" refers to the number of county-level changes in coverage that are used in estimation. The test for pre-trends is computed by conducting a Wald test using the bootstrapped covariance matrix. For each panel, I exclude from the sample all ever-treated and never-treated counties with fewer than 100 tweets in the relevant topic area over the entire sample period. Error bars represent pointwise 95 percent confidence intervals. Standard errors are clustered at the county level.

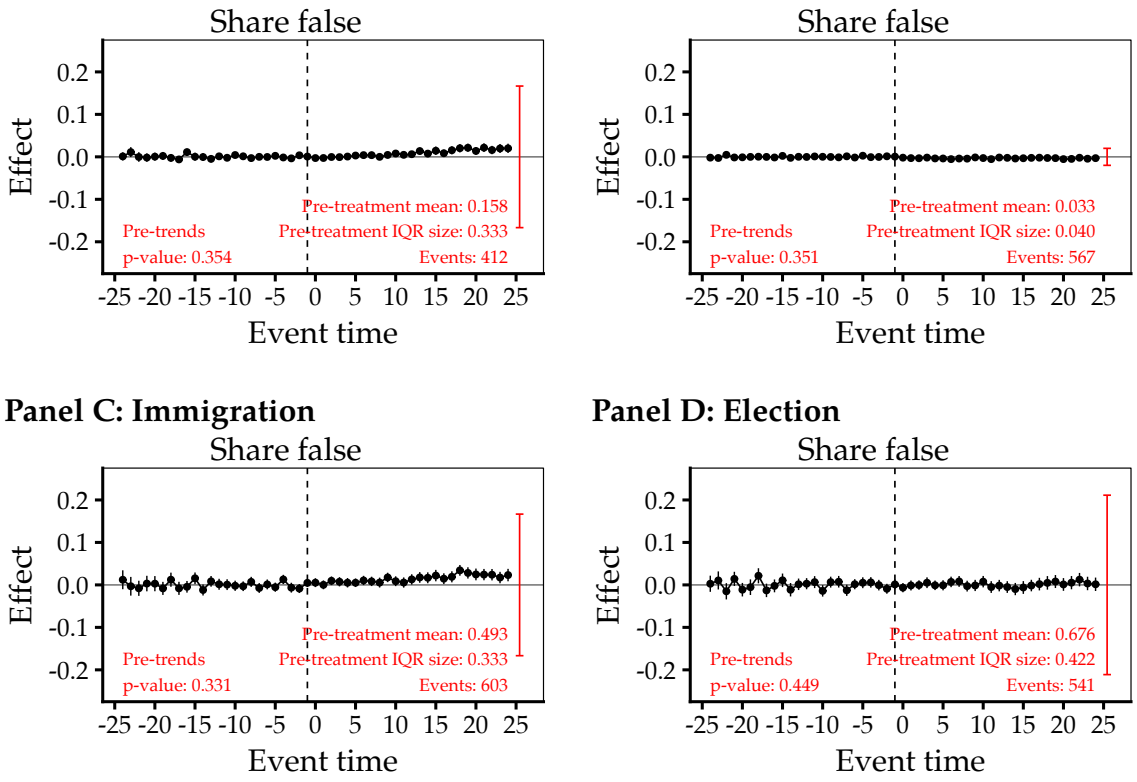
tion (panel C). I find slight negative effects of a coverage change on the share of false tweets about climate change (panel B). I do not detect effects on the share of false tweets about U.S. election results (panel D). All panels fail to reject the null hypothesis of no pre-trends.

Each plot provides several annotations in red to contextualize the economic significance of the treatment effects. For instance, I compute that for all ever-treated counties in all months prior to treatment, the mean share of vaccine-related tweets that contain misinformation is 0.158, and the mean share of immigration-related tweets that contain misinformation is 0.493. In addition, I compute that the distance between the 25th percentile and 75th percentile of the outcome variable is 0.333 for both vaccines and for immigration. I show the magnitude of the interquartile range on the right side of each plot to contextualize the economic significance of the treatment effects. For ease of visual comparison, I center this "ruler" at zero. Note that the visualization does not represent the actual 25th or 75th percentile values of the outcome variable, nor does it indicate where the interquartile range lies in levels.

These event study aggregations have a key drawback: All treatment dosages are considered the same. At each time horizon, units that experience different types of treatment — say, a weekly newspaper opening in one county but closing in another — are combined together. To resolve this problem, I also present *normalized* event study aggregations in figure 13. This figure provides the same treatment effects as those in figure 12, only this time, prior to aggregation, each $ATT(g,t,d)$ cell is divided by its dosage prior to aggregation. These estimates can be interpreted as the average *per-dosage* treatment effect for ever-treated units in each period relative to treatment.

Table 6 presents simple aggregates of ATTs for each topic in order to provide pointwise estimates over the entire 24-month post-treatment period. Column 1 provides estimates without normalization. The average treatment effect for ever-treated counties was 2.3 percentage points on the share of vaccine-related tweets that contain misinformation (15 percent of the pre-treatment mean and 7 percent of the IQR), and 3.8 percentage points on the share of immigration-related tweets that contain misinformation (8 percent of the pre-treatment mean and 11 percent of the IQR). These magnitudes can be further contextualized given that the aver-

Figure 13: Normalized event study aggregates of ATTs, share of tweets containing misinformation



Notes: This figure plots normalized "event study" aggregates of ATTs for the 24 months (2 years) before and after a county-level coverage change. The ATT parameter and aggregation procedure is described in detail in section 5 of the text. The dependent variable is the share of tweets posted in a given county in a given month that the relevant supervised classifier labels as likely to contain misinformation. Each panel provides normalized event study estimates for one of the four topic areas I have identified as prone to misinformation. Each $ATT(g,t,d)$ cell is divided by the dosage prior to aggregation. Annotations are provided in red to contextualize treatment effects for each figure. I pool the dependent variable for all ever-treated counties across all pre-treatment periods and report both the mean and difference between the 25th percentile value and the 75th percentile value. The size of the interquartile range is plotted, centered at zero, on the right side of each figure. The number of "events" refers to the number of county-level changes in coverage that are used in estimation. The test for pre-trends is computed by conducting a Wald test using the bootstrapped covariance matrix. For each panel, I exclude from the sample all ever-treated and never-treated counties with fewer than 100 tweets in the relevant topic area over the entire sample period. Error bars represent pointwise 95 percent confidence intervals. Standard errors are clustered at the county level.

age treatment dosage is approximately 2.58. In other words, the average coverage change consists of a decrease in publication days per week of 2.58. This treatment dosage is roughly equivalent to 1.5 weekly newspapers going out of business.

Column 2 provides normalized estimates. For a one-unit change in dosage, the average treatment effect for ever-treated counties was about 1 percentage point on the share of vaccine-related tweets that contain misinformation (6 percent of the pre-treatment mean and 3 percent of the IQR), and 1.5 percentage points on the share of immigration-related tweets that contain misinformation (3 percent of the pre-treatment mean and 4 percent of the IQR).

Table 6: Simple aggregates of ATTs, share of tweets containing misinformation

Treatment effect	All counties	All counties normalized	Republican counties	Democratic counties	Dosage response	$\beta^{Rep} = \beta^{Dem}$ p-value
Vaccine	0.023	0.009	0.013	0.027	0.012	0.285
	(0.006)	(0.002)	(0.007)	(0.010)	(0.002)	
Climate	-0.009	-0.003	-0.008	-0.002	-0.000	0.432
	(0.004)	(0.002)	(0.006)	(0.006)	(0.001)	
Immigration	0.038	0.015	0.025	0.057	0.013	0.162
	(0.011)	(0.004)	(0.014)	(0.018)	(0.004)	
Election	0.002	0.001	0.001	0.020	0.000	0.489
	(0.012)	(0.005)	(0.014)	(0.023)	(0.004)	

Notes: This table provides simple aggregates of ATTs for the 24 months (2 years) before and after a county-level coverage change. The ATT parameter and aggregation procedure is described in detail in section 5 of the text. The dependent variable is the share of tweets posted in a given county in a given month that the relevant supervised classifier labels as likely to contain misinformation. Column 1 presents aggregated treatment effects for all ever-treated counties. Corresponding event study aggregations can be found in figure 12. Column 2 reports the same aggregates, except each $ATT(g, t, d)$ cell is divided by the dosage prior to aggregation. Corresponding event study aggregations can be found in figure 13. Column 3 presents aggregated treatment effects for all ever-treated counties labeled Rep. Column 4 presents aggregated treatment effects for all ever-treated counties labeled Dem. The labeling process is described in section 6.6 of the text. Corresponding event study aggregations can be found in figure 15. Column 5 reports the slope of the line of best fit through the dosage response estimates provided in figure 14. Column 6 reports p-values for an asymptotic hypothesis test that the coefficients in columns 3 and 4 are equal to each other. I exclude from the sample all ever-treated and never-treated counties with fewer than 100 tweets in the relevant topic area over the entire sample period. Standard errors are clustered at the county level.

The estimates in column 2 of table 6 are most appropriate to compare to the naive cross-sectional estimates in panel B of table 4. The normalized treatment effects measure the per-dosage effect of treatment over the entire post-treatment

period. The cross-sectional estimates measure the effects of a one-unit change in the coverage index, which is equivalent to a dosage of -1. Notice that the magnitudes of the normalized treatment effects are less than half the size of the naive cross-sectional estimates presented in table 4.

I am somewhat puzzled by the slight negative effects on the composition of climate change-related tweets. In general, I am less confident in my results for climate change-related outcomes and election-related outcomes, because, as explained in section 3.2, the supervised classifiers are less likely to generalize well to the tweets in the Geotweet Archive. Consequently, the predicted labels may contain more noise in those two topic areas than in the vaccines and immigration topic areas. In particular, very few tweets about climate change are labeled as false to begin with. For all ever-treated counties in all months prior to treatment, the mean share of climate-related tweets that contain misinformation is just 0.033. Consequently, I am not confident that the effect I am picking up is distinguishable from noise.

6.5 Dosage-response effects

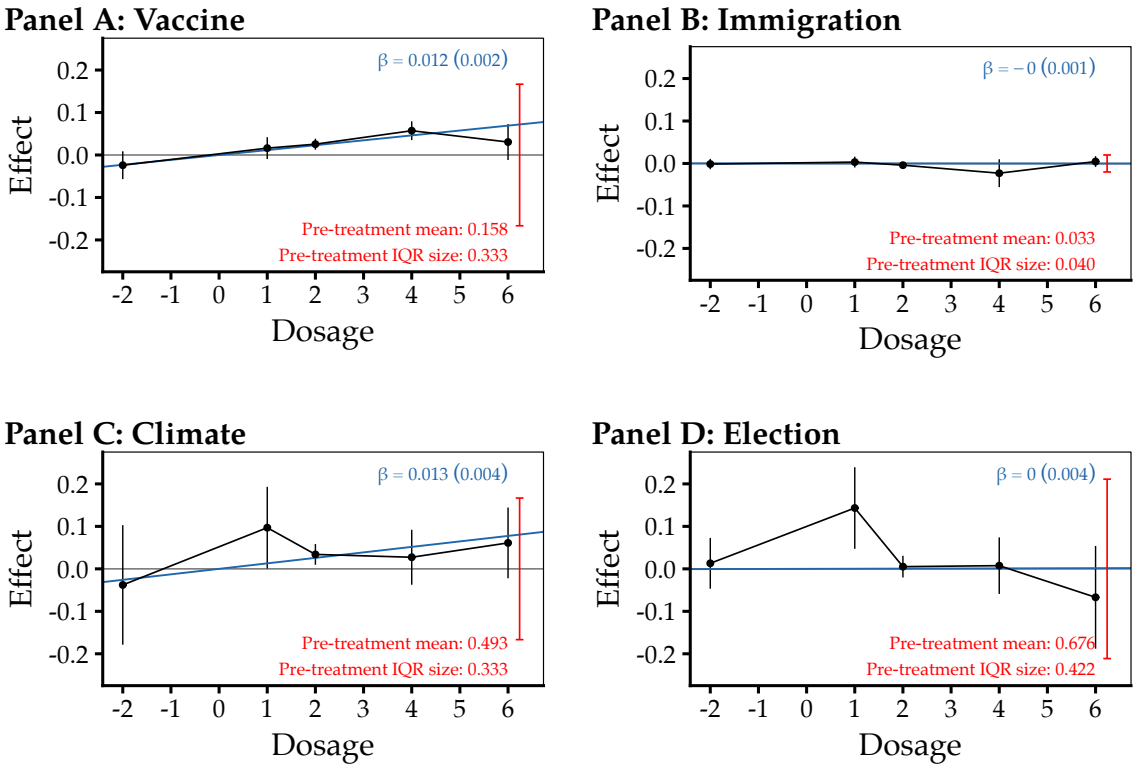
Figure 14 plots dosage-response aggregates of ATTs. Each panel presents treatment effects on the composition of Twitter activity for a given topic area. Each dot represents the average treatment effect on ever-treated counties that experienced a coverage change of a given dosage, pooled over all post-treatment periods.

I only plot dosage-specific estimates for dosages that have sufficient identifying variation. If a given dosage is only observed for a limited number of counties in a limited number of groups, I do not have sufficient identifying variation to estimate precise dosage-specific ATTs. Such estimates risk over-interpreting noise driven by a small number of units. Accordingly, I restrict the figure to only plot dosage-specific ATTs for dosages with adequate cross-county and cross-group variation. Specifically, I require that a given dosage be observed in at least 10 distinct groups.

The five dosages that remain are the five most common types of coverage changes. Each coverage change can (roughly) be interpreted as follows:

- $d = -2$: A new weekly newspaper opens.
- $d = 1$: A newspaper reduces publication frequency by one day per week.

Figure 14: Dosage-response aggregates of ATTs, share of tweets containing misinformation



Notes: This figure plots "dosage-response" aggregates of ATTs for the 24 months (2 years) before and after a county-level coverage change. The ATT parameter and aggregation procedure is described in detail in section 5 of the text. The dependent variable is the share of tweets posted in a given county in a given month that the relevant supervised classifier labels as likely to contain misinformation. Each panel provides dosage-response estimates for one of the four topic areas I have identified as prone to misinformation. Only dosage-specific estimates that contain sufficient group-level identifying variation are included. A positive dosage indicates that a county has lost newspaper coverage. A negative dosage indicates that a county has gained newspaper coverage. The line of best fit is plotted in blue. It is subject to the constraint that it must intersect zero when dosage is zero. It is calculated using weighted least squares, where the weights are inversely proportional to the variance of the estimated treatment effect for each dosage. Annotations are provided in red to contextualize treatment effects for each figure. I pool the dependent variable for all ever-treated counties across all pre-treatment periods and report both the mean and difference between the 25th percentile value and the 75th percentile value. The size of the interquartile range is plotted, centered at zero, on the right side of each figure. Annotations are provided in blue to contextualize the slope of the line of best fit. For each panel, I exclude from the sample all ever-treated and never-treated counties with fewer than 100 tweets in the relevant topic area over the entire sample period. Error bars represent pointwise 95 percent confidence intervals. Standard errors for the treatment effects are clustered at the county level.

- $d = 2$: An existing weekly newspaper shuts down.
- $d = 4$: A daily newspaper becomes a weekly newspaper.
- $d = 6$: A daily newspaper shuts down.

In each figure, the blue line is the line of best fit for the treatment effect sizes, subject to the constraint that the line must intersect zero when the treatment dosage is zero. I weight treatment effects inversely proportional to the size of their standard errors. The slope of this line tests whether there is a proportional relationship between treatment dosage and the size of the treatment effect.

As with the event study aggregations, annotations are provided to contextualize treatment effect magnitudes. The statistics for the pre-treatment mean and the magnitude of the interquartile range are the same as the ones provided with the plots of the event study aggregations in figure 12.

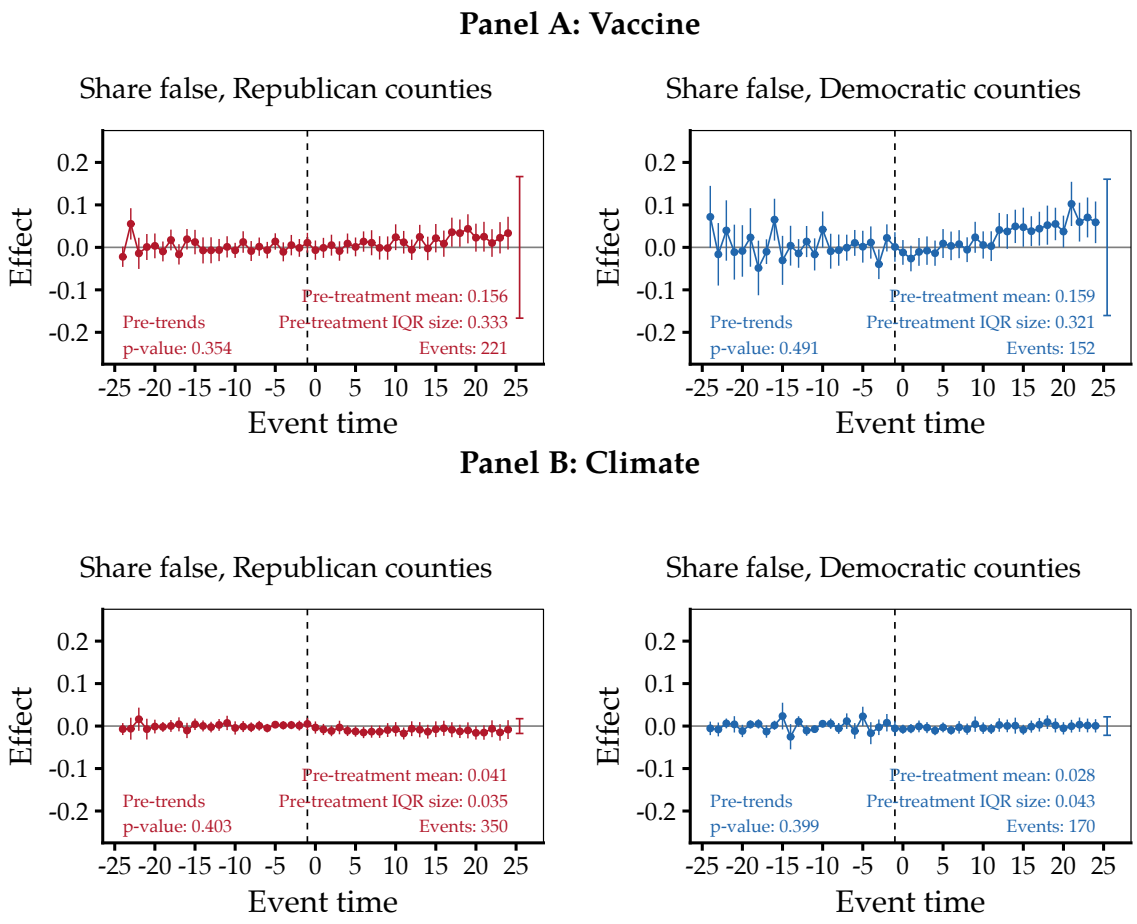
Column 4 of table 6 provides estimates for the slope of this line for each topic area. For vaccines and immigration, the size of the treatment effect appears to be directly proportional to the dosage. A one unit increase in the dosage increases the treatment effect size on the share of false vaccine-related tweets and on the share of false immigration-related tweets by approximately 1.2 percentage points and 1.3 percentage points respectively. As mentioned in section 6.4, this treatment effect is less than half the size of the estimate calculated using the naive cross-sectional approach in section 6.2.

6.6 Treatment effects by partisanship

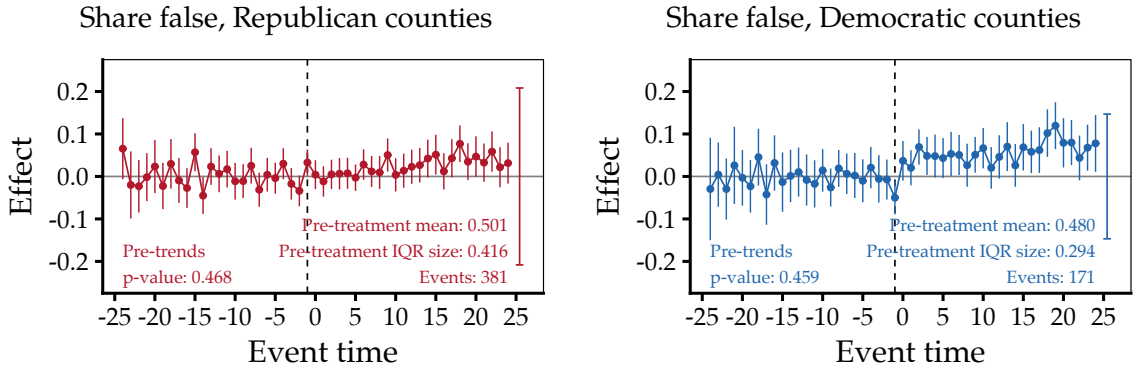
Next, figure 15 examines whether there is heterogeneity in treatment effects on the composition of Twitter discourse based on the political partisanship of the treated county. I assign a label of Dem or Rep to each county based on the presidential election results for that county over the duration of the entire sample period. If a plurality of voters in a given county votes for the Republican (Democratic) candidate in three of the four presidential elections in 2008, 2012, 2016, and 2020, I label the county as Rep (Dem).

The left figure of each panel estimates treatment effects on ever-treated counties that were assigned the Rep label. Comparison counties are never-treated counties

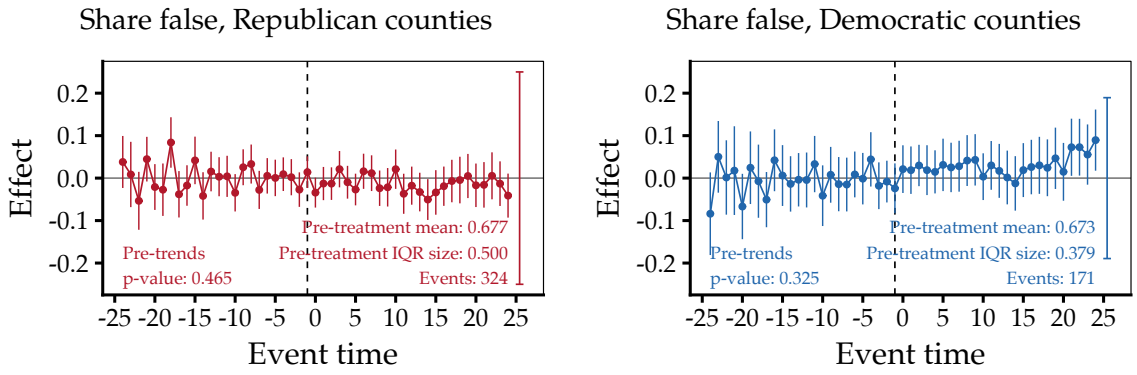
Figure 15: Event study aggregates of ATTs, share of tweets containing misinformation by partisanship



Panel C: Immigration



Panel D: Election



Notes: This figure plots "event study" aggregates of ATTs for the 24 months (2 years) before and after a county-level coverage change. The ATT parameter and aggregation procedure is described in detail in section 5 of the text. The dependent variable is the share of tweets posted in a given county in a given month that the relevant supervised classifier labels as likely to contain misinformation. Each panel provides event study estimates for one of the four topic areas I have identified as prone to misinformation. The left figure plots treatment effects for ever-treated counties labeled as Rep. The right figure plots treatment effects for ever-treated counties labeled as Dem. The labeling process is described in section 6.6. Annotations are provided to contextualize treatment effects for each figure. I pool the dependent variable for all ever-treated counties across all pre-treatment periods and report both the mean and difference between the 25th percentile value and the 75th percentile value. The size of the interquartile range is plotted, centered at zero, on the right side of each figure. The number of "events" refers to the number of county-level changes in coverage that are used in estimation. The test for pre-trends is computed by conducting a Wald test using the bootstrapped covariance matrix. For each panel, I exclude from the sample all ever-treated and never-treated counties with fewer than 100 tweets in the relevant topic area over the entire sample period. Error bars represent pointwise 95 percent confidence intervals. Standard errors are clustered at the county level.

that were also assigned the Rep label. The right figure of each panel estimates treatment effects on ever-treated units that were assigned the Dem label. Comparison counties are never-treated counties that were also assigned the Dem label.

For both Republican and Democratic counties, dynamic treatment effects are statistically significant and meaningful. As with the event study aggregations in figure 12, annotations are provided to contextualize treatment effect magnitudes. In addition, I fail to reject the null hypothesis of no pre-trends.

Columns 3 and 4 of table 6 present simple aggregates of ATTs for each topic by partisanship for the entire 24-month post-treatment period. The average treatment effect on the composition of vaccine-related Twitter discourse for ever-treated Democratic counties (2.7 percentage points) is almost double the average treatment effect for Republican counties (1.5 percentage points). Likewise, the average treatment effect on the composition of immigration-related Twitter discourse for ever-treated Democratic counties (5.4 percentage points) is more than double the average treatment effect for Republican counties (2.3 percentage points).

Column 6 of table 6 reports p-values for hypotheses that test whether, for each topic area, the average treatment effect on treated Democratic counties is equal to the average treatment effect on treated Republican counties. I cannot reject the null that both treatment effects are the same at a 95 percent confidence level for any topic area.

7 Conclusion

This paper investigates how the supply of trustworthy local news shapes the information environment on social media. Using a panel of 2.7 billion geotagged tweets and a novel dataset tracking local newspaper coverage changes across U.S. counties from 2012 to 2023, I find that declines in local newspaper coverage leave the overall volume of Twitter discourse essentially unchanged, but meaningfully increase the share of discourse that contains misinformation. In counties that experience a coverage change, the share of vaccine-related tweets containing misinformation rises by an average of 2.3 percentage points over the subsequent two years, and the share of immigration-related tweets containing misinformation rises by 3.8 percentage

points. These effects are substantial relative to pre-treatment means. In addition, estimated treatment effects appear to be proportional to the intensity of treatment for both vaccine-related misinformation and immigration-related misinformation. Furthermore, effect sizes are about twice as large for Democratic counties as they are for Republican counties, although Republican counties tend to have a higher baseline share of tweets that contain misinformation. Taken together, the results suggest that local newspapers do not merely serve as sources of information, but rather as anchors of an information ecosystem that suppresses the salience of false claims on social media.

These findings have important implications for ongoing policy debates about the future of local journalism. The decline of local newspapers over the past two decades is well-documented, and the results presented here suggest that this decline has externalities that extend beyond the domains — such as civic engagement and electoral accountability — that are traditionally associated with local news. Rather, the availability of local news seems to improve the quality of online discourse itself.

Although no single intervention is likely to reverse the structural economic forces driving newspaper closures, policies that subsidize local news production — through tax incentives, public funding, or philanthropic support — may generate social returns that extend beyond the direct readership of those publications. More broadly, this paper highlights the importance of the information supply chain: The veracity of what circulates on social media is not determined solely by the platforms themselves, but also by the availability of credible alternatives that compete for readers' attention.

Future research should further explore the mechanisms by which changes to the supply of information affect the users' consumption and production of content on social media. One possibility is that the loss of trustworthy information induces individuals who previously relied on print or online local news to migrate toward lower-quality information sources, including social media. A second possibility is that changes to the composition of online information primarily occur through the users that are already consuming news from both the trustworthy source and the untrustworthy source. Then, shocks to the information supply may affect these

users' abilities to handle the narratives they encounter online. These two mechanisms are not mutually exclusive, and distinguishing between them empirically is difficult. However, understanding the dominant channel is important for policymakers, because they point to different policy levers. If it is the former, then media literacy interventions — such as the ones proposed by the "pre-bunking" literature — targeting new social media users may have an outsized effect in addressing the proliferation of online misinformation. If it is the latter, though, then policy should focus on ensuring that users with demand for news have trustworthy options that are easily accessible.

References

Acemoglu, D., A. Ozdaglar, and J. Siderius (2024). A model of online misinformation. *Review of Economic Studies 91*(6), 3117–3150.

Acemoglu, D., A. Ozdaglar, and J. Siderius (2025). AI and social media: A political economy perspective. Working Paper 33892, National Bureau of Economic Research.

Ahmad, W., A. Sen, C. Eesley, and E. Brynjolfsson (2024). The role of advertisers and platforms in monetizing misinformation: Descriptive and experimental evidence. Working Paper 32187, National Bureau of Economic Research.

Akesson, J., R. W. Hahn, R. D. Metcalfe, and M. Monti-Nussbaum (2023). The impact of fake reviews on demand and welfare. Working Paper 31836, National Bureau of Economic Research.

Allcott, H., L. Braghieri, S. Eichmeyer, and M. Gentzkow (2020). The welfare effects of social media. *American Economic Review 110*(3), 629–676.

Allcott, H. and M. Gentzkow (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives 31*(2), 211–236.

Allcott, H., M. Gentzkow, R. Levy, A. Crespo-Tenorio, N. Dumas, W. Mason, D. Moehler, P. Barbera, T. W. Brown, J. C. Cisneros, D. Dimmery, D. Freelon, S. González-Bailón, A. M. Guess, Y. M. Kim, D. Lazer, N. Malhotra, S. Nair-Desai, B. Nyhan, A. C. P. de Queiroz, J. Pan, J. Settle, E. Thorson, R. Tromble, C. V. Rivera, B. Wittenbrink, M. Wojcieszak, S. Yang, S. Zahedian, A. Franco, C. K. de Jonge, N. J. Stroud, and J. A. Tucker (2025). The effects of political advertising on Facebook and Instagram before the 2020 US election. Working Paper 33818, National Bureau of Economic Research.

Allcott, H., M. Gentzkow, and C. Yu (2019). Trends in the diffusion of misinformation on social media. *Research & Politics 6*(2), 2053168019848554.

Almond, D., X. Du, and A. Vogel (2020). Russian holidays predict troll activity 2015-2017. Working Paper 28035, National Bureau of Economic Research.

Alnabhan, M. Q. and P. Branco (2024). Fake news detection using deep learning: A systematic literature review. *IEEE Access 12*, 114435–114459.

Angelucci, C., J. Cagé, and M. Sinkinson (2024). Media competition and news diets. *American Economic Journal: Microeconomics 16*(2), 62–102.

Angelucci, C., M. Gutmann, and A. Prat (2024). Beliefs about political news in the run-up to an election. Working Paper 32802, National Bureau of Economic Research.

Asirvatham, H., E. Mokski, and A. Shleifer (2026). GPT as a measurement tool. Working Paper 34834, National Bureau of Economic Research.

Athey, S., E. Calvano, and J. Gans (2013). The impact of the internet on advertising markets for news media. Working Paper 19419, National Bureau of Economic Research.

Athey, S., K. Grabarz, M. Luca, and N. C. Wernerfelt (2023). Digital public health interventions at scale: The impact of social media advertising on beliefs and outcomes related to COVID vaccines. *Proceedings of the National Academy of Sciences 120*(5), e2208110120.

Athey, S., M. Mobius, and J. Pal (2021). The impact of aggregators on internet news consumption. Working Paper 28746, National Bureau of Economic Research.

Azzimonti, M. and M. Fernandes (2023). Social media networks, fake news, and polarization. *European Journal of Political Economy 76*, 102256.

Bang, M., L. LḾeude, A. Postlewaite, and H. Sieg (2023). Access and exposure to local news media in the digital era: Evidence from U.S. media markets. Working Paper 31436, National Bureau of Economic Research.

Beach, B. and W. W. Hanlon (2023). Historical newspaper data: A researcher's guide. *Explorations in Economic History 90*, 101541.

Beattie, G., R. Durante, B. Knight, and A. Sen (2021). Advertising spending and media bias: Evidence from news coverage of car safety recalls. *Management Science 67*(2), 698–719.

Bhuller, M., T. Havnes, J. McCauley, and M. Mogstad (2024). How the internet changed the market for print media. *American Economic Journal: Applied Economics 16*(2), 318–358.

Bowen, R., D. Dmitriev, and S. Galperti (2023). Learning from shared news: When abundant information leads to belief polarization. *The Quarterly Journal of Economics 138*(2), 955–1000.

Bursztyn, L., G. Egorov, R. Enikolopov, and M. Petrova (2019). Social media and xenophobia: Evidence from Russia. Working Paper 26567, National Bureau of Economic Research.

Bursztyn, L., B. R. Handel, R. Jimenez, and C. Roth (2023). When product markets become collective traps: The case of social media. Working Paper 31771, National Bureau of Economic Research.

Bursztyn, L., A. Rao, C. Roth, and D. Yanagizawa-Drott (2023). Opinions as facts. *Review of Economic Studies 90*(4), 1832–1864.

Callaway, B., A. Goodman-Bacon, and P. H. C. Sant'Anna (2024). Difference-in-differences with a continuous treatment. Working Paper 32117, National Bureau of Economic Research.

Callaway, B. and P. H. C. Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics 225*(2), 200–230.

Campante, F., R. Durante, F. Hagemeister, and A. Sen (2025). GenAI misinformation, trust, and news consumption: Evidence from a field experiment. Working Paper 34100, National Bureau of Economic Research.

Cao, A., J. M. Lindo, and J. Zhong (2023). Can social media rhetoric incite hate incidents? Evidence from Trump's "Chinese Virus" tweets. *Journal of Urban Economics 137*, 103590.

Cao, J., H. H. Hochmair, and F. Basheeh (2022). The effect of twitter app policy changes on the sharing of spatial information through twitter users. *Geographies 2*(3), 549–562.

Casillas, A., M. Farboodi, L. Hashemi, M. Saeedi, and S. Wilson (2024). (dis)information wars. Working Paper 32896, National Bureau of Economic Research.

Castillo, C., M. Mendoza, and B. Poblete (2011). Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, New York, NY, USA, pp. 675—684. Association for Computing Machinery.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng (2024). On binscatter. *American Economic Review 114*(5), 1488–1514.

Center for Innovation and Sustainability in Local Media (2024). US news deserts database archive [dataset].

Chen, J. and J. Roth (2024). Logs with zeros? Some problems and solutions. *Quarterly Journal of Economics 139*(2), 891–936.

Chen, Y., H. Fang, Y. Zhao, and Z. Zhao (2024). Recovering overlooked information in categorical variables with LLMs: An application to labor market mismatch. Working Paper 32327, National Bureau of Economic Research.

Chiou, L. and C. Tucker (2017). Content aggregation by platforms: The case of the news media. *Journal of Economics & Management Strategy 26*(4), 782–805.

Chiou, L. and C. Tucker (2018). Fake news and advertising on social media: A study of the Anti-Vaccination movement. Working Paper 25223, National Bureau of Economic Research.

Compton, J., S. van der Linden, J. Cook, and M. Basol (2021). Inoculation theory in the post-truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories. *Social and Personality Psychology Compass 15*(6), e12602.

CQ Press (2024). CQ voting and elections collection [dataset].

Dadkhah, S., X. Zhang, A. G. Weismann, A. Firouzi, and A. A. Ghorbani (2023, oct). The largest social media ground-truth dataset for real/fake content: Truthseeker. *IEEE Transactions on Computational Social Systems 99*, 1–15.

de Chaisemartin, C. and X. D'Haultfoeuille (2023). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *The Econometrics Journal 26*(3), C1–C30.

DellaVigna, S. and E. Kaplan (2007). The Fox News effect: Media bias and voting. *The Quarterly Journal of Economics 122*(3), 1187–1234.

Di Tella, R. and I. Franceschelli (2011). Government advertising and media coverage of corruption scandals. *American Economic Journal: Applied Economics 3*(4), 119–151.

Diggelmann, T., J. Boyd-Graber, J. Bulian, M. Ciaramita, and M. Leippold (2020, dec). Climate-fever: A dataset for verification of real-world climate claims. In *Tackling Climate Change with Machine Learning Workshop at NeurIPS 2020*, Online. Online, 11 December 2020.

Dyck, A., N. Volchkova, and L. Zingales (2008). The corporate governance role of the media: Evidence from Russia. *The Journal of Finance 63*(3), 1093–1135.

Ewens, M., A. Gupta, and S. T. Howell (2022). Local journalism under private equity ownership. Working Paper 29743, National Bureau of Economic Research.

Ferrara, A., J. Y. Ha, and R. Walsh (2022). Using digitized newspapers to refine historical measures: The case of the Boll Weevil. Working Paper 29808, National Bureau of Economic Research.

Filippas, A., J. J. Horton, E. Lipnowski, and P. Parasurama (2021). The production and consumption of social media. Working Paper 28666, National Bureau of Economic Research.

Freyaldenhoven, S., C. Hansen, J. Pérez Pérez, and J. M. Shapiro (2021). Visualization, identification, and estimation in the linear panel event-study design. Working Paper 29170, National Bureau of Economic Research.

Fujiwara, T., K. Müller, and C. Schwarz (2024). The effect of social media on elections: Evidence from the United States. *Journal of the European Economic Association 22*(3), 1495–1539.

Gandhi, A., B. Hollenbeck, and Z. Li (2025). Misinformation and mistrust: The equilibrium effects of fake reviews on Amazon.com. Working Paper 34161, National Bureau of Economic Research.

Ganin, Y., E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research 17*(1), 1–35.

Genesove, D. (1999). The adoption of offset presses in the daily newspaper industry in the United States. Working Paper 7076, National Bureau of Economic Research.

Gentzkow, M. (2007). Valuing new goods in a model with complementarity: Online newspapers. *American Economic Review 97*(3), 713–744.

Gentzkow, M., E. L. Glaeser, and C. Goldin (2004). The rise of the fourth estate: How newspapers became informative and why it mattered. Working Paper 10791, National Bureau of Economic Research.

Gentzkow, M., N. Petek, J. M. Shapiro, and M. Sinkinson (2015). Do newspapers serve the state? Incumbent party influence on the U.S. press, 1869–1928. *Journal of the European Economic Association 13*(1), 29–61.

Gentzkow, M. and J. M. Shapiro (2010). What drives media slant? Evidence from U.S. daily newspapers. *Econometrica 78*(1), 35–71.

Gentzkow, M. and J. M. Shapiro (2011). Ideological segregation online and offline. *The Quarterly Journal of Economics 126*(4), 1799–1839.

Gentzkow, M. and J. M. Shapiro (2013). Ideology and online news. Working Paper 19675, National Bureau of Economic Research.

Gentzkow, M., J. M. Shapiro, and M. Sinkinson (2011). The effect of newspaper entry and exit on electoral politics. *American Economic Review 101*(7), 2980–3018.

Gentzkow, M., J. M. Shapiro, and M. Sinkinson (2014). Competition and ideological diversity: Historical evidence from U.S. newspapers. *American Economic Review 104*(10), 3073–3114.

George, L. and J. Waldfogel (2003). Who affects whom in daily newspaper markets? *Journal of Political Economy 111*(4), 765–784.

Glaeser, E. L. and G. Ujhelyi (2010). Regulating misinformation. *Journal of Public Economics 94*(3-4), 247–257.

Grossman, G. M. and E. Helpman (2023). Electoral competition with fake news. *European Journal of Political Economy 77*, 102315.

Harjani, T., J. Roozenbeek, M. Biddlestone, S. van der Linden, A. Stuart, M. Iwahara, B. Piri, R. Xu, B. Goldberg, and M. Graham (2022). A practical guide to prebunking misinformation. Technical report, University of Cambridge and BBC Media Action and Jigsaw (Google).

Hayawi, K., S. Shahriar, M. A. Serhani, I. Taleb, and S. S. Mathew (2022). Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. *Public Health 203*, 23–30.

Heese, J. and J. Pacelli (2024). The monitoring role of social media. *Review of Accounting Studies 29*(2), 1666–1706.

Heese, J., G. Pérez-Cavazos, and C. D. Peter (2022). When the local newspaper leaves town: The effects of local newspaper closures on corporate misconduct. *Journal of Financial Economics 145*(2), 445–463.

Ho, L. Y., E. Breza, M. Alsan, A. Banerjee, A. G. Chandrasekhar, F. C. Stanford, R. Fior, P. Goldsmith-Pinkham, K. Holland, E. Hoppe, L.-M. Jean, L. Ogbu-Nwobodo, B. A. Olken, C. Torres, P.-L. Vautrey, E. Warner, and E. Duflo (2023). The impact of large-scale social media advertising campaigns on COVID-19 vaccination: Evidence from two randomized controlled trials. *AEA Papers and Proceedings 113*, 653–658.

Hossain, S., A. Mladenovic, Y. Chen, and G. Gidel (2024, jul). A persuasive approach to combating misinformation. In *Proceedings of the 41st International Conference on Machine Learning*, Volume 235 of *Proceedings of Machine Learning Research*, pp. 18926–18943. PMLR.

Kaliyar, R. K., A. Goswami, and P. Narang (2021). Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia Tools and Applications 80*(8), 11765–11788.

Kartal, M. and J.-R. Tyran (2022). Fake news, voter overconfidence, and the quality of democratic choice. *American Economic Review 112*(10), 3367–3397.

Knight, B. G. and C.-F. Chiang (2011). Media bias and influence: Evidence from newspaper endorsements. *The Review of Economic Studies 78*(3), 795–820.

Kominers, S. D. and J. M. Shapiro (2025). Robust content moderation: Theory and applications. Working Paper 32156, National Bureau of Economic Research.

Larcinese, V., R. Puglisi, J. M. Snyder, and Jr. (2011). Partisan bias in economic news: Evidence on the agenda-setting behavior of U.S. newspapers. *Journal of Public Economics 95*(9-10), 1178–1189.

Levy, R. (2021, March). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review 111*(3), 831–70.

Lewandowsky, S. and S. van der Linden (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology 32*(2), 348–384.

Lewis, B. and D. Jain (2016). Harvard CGA Geotweet archive v2.0 [dataset].

List, J. A., L. M. Ramírez, J. Seither, J. Unda, and B. Vallejo (2024). Toward an understanding of the economics of misinformation: Evidence from a demand side field experiment on critical thinking. Working Paper 32367, National Bureau of Economic Research.

Ludwig, J., S. Mullainathan, and A. Rambachan (2025). Large language models: An applied econometric framework. Working Paper 33344, National Bureau of Economic Research.

Ma, J., W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha (2016, jul). Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, New York, New York, USA, pp. 3818–3824. AAAI Press.

Mullainathan, S. and A. Shleifer (2005). The market for news. *American Economic Review 95*(4), 1031–1053.

Müller, K. and C. Schwarz (2023). From hashtag to hate crime: Twitter and antiminority sentiment. *American Economic Journal: Applied Economics 15*(3), 270–312.

Nyhan, B. (2020, August). Facts and myths about misperceptions. *Journal of Economic Perspectives 34*(3), 220–36.

Oberholzer-Gee, F. and J. Waldfogel (2009). Media markets and localism: Does local news en Español boost hispanic voter turnout? *American Economic Review 99*(5), 2120–2128.

Pennycook, G., Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, and D. G. Rand (2021). Shifting attention to accuracy can reduce misinformation online. *Nature 592*(7855), 590–595.

Pennycook, G., J. McPhetres, Y. Zhang, J. G. Lu, and D. G. Rand (2020). Fighting Covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science 31*(7), 770–780.

Pew Research Center (2025, oct). How americans' trust in information from news organizations and social media sites has changed over time.

Puglisi, R., J. M. Snyder, and Jr. (2011). Newspaper coverage of political scandals. *The Journal of Politics 73*(3), 931–950.

Puglisi, R., J. M. Snyder, and Jr. (2015). The balanced U.S. press. *Journal of the European Economic Association 13*(2), 240–264.

Qian, N. and D. Yanagizawa-Drott (2017). Government distortion in independently owned media: Evidence from U.S. news coverage of human rights. *Journal of the European Economic Association 15*(2), 463–499.

Rashkin, H., E. Choi, J. Y. Jang, S. Volkova, and Y. Choi (2017, sep). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 2931–2937. Association for Computational Linguistics.

Roozenbeek, J. and S. van der Linden (2024). *The Psychology of Misinformation*. Cambridge University Press.

Roth, J., P. H. C. Sant'Anna, A. Bilinski, and J. Poe (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics 235*(2), 2218–2244.

Sarkar, S. K. (2025). Economic representations. Technical report, Working Paper.

Schulhofer-Wohl, S. and M. Garrido (2013). Do newspapers matter? Short-Run and Long-Run evidence from the closure of The Cincinnati Post. *Journal of Media Economics 26*(2), 60–81.

Snyder, J. M., Jr., and D. Strömberg (2010). Press coverage and political accountability. *Journal of Political Economy 118*(2), 355–408.

Society of Professional Journalists (2014, sep). SPJ code of ethics. Revised September 6, 2014.

Thibault, C., J.-J. Tian, G. Péloquin-Skulski, T. L. Curtis, J. Zhou, F. Laflamme, L. Y. Guan, R. Rabbany, J.-F. Godbout, and K. Pelrine (2025). A guide to misinformation detection data and evaluation. KDD '25, New York, NY, USA, pp. 5801–5809. Association for Computing Machinery.

Thorne, J., A. Vlachos, C. Christodoulopoulos, and A. Mittal (2018, jun). Fever: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana, pp. 809–819. Association for Computational Linguistics.

U.S. Census Bureau (2025). TIGER/Line shapefiles for U.S. counties [dataset].

van Binsbergen, J. H., S. Bryzgalova, M. Mukhopadhyay, and V. Sharma (2024). (almost) 200 years of news-based economic sentiment. Working Paper 32026, National Bureau of Economic Research.

van der Linden, S. (2024). Countering misinformation through psychological inoculation. In *Advances in Experimental Social Psychology*, pp. 1–58. Elsevier.

van der Linden, S., D. Louison-Lavoy, N. Blazer, N. S. Noble, and J. Roozenbeek (2026, jan). Prebunking misinformation techniques in social media feeds: Results from an instagram field study. *Harvard Kennedy School Misinformation Review*.

Vosoughi, S., D. Roy, and S. Aral (2018). The spread of true and false news online. *Science 359*(6380), 1146–1151.

Wang, T. (2023). The electric telegraph, news coverage and political participation. Working Paper 31468, National Bureau of Economic Research.

Wang, W. Y. (2017, jul). "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada, pp. 422–426. Association for Computational Linguistics.

Wooldridge, J. M. (2023). Simple approaches to nonlinear difference-in-differences with panel data. *The Econometrics Journal 26*(3), C31–C66.

Xu, W., J. Wu, Q. Liu, S. Wu, and L. Wang (2022). Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM Web Conference 2022*, WWW '22, New York, NY, USA, pp. 2501—2510. Association for Computing Machinery.

Yang, Y.-T., T. Li, and Q. Zhu (2023). Designing policies for truth: Combating misinformation with transparency and information design. In *2023 21st International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, pp. 127–134. IEEE.

A Classification of tweets

A.1 Topic classification schema

In order to classify the tweets in the Geotweet Archive into topic areas, I prompted the Qwen3 mixture-of-experts model using the following schema:

```
### Instruction
Please provide topic labels for the user query.
   Select only the most relevant topic labels. If no
   label is obviously relevant, select "other". Only
   include labels listed in all_topic_labels.
### User Query
<<QUERY>>
### Topic labels
all_topic_labels = [
    "climate_change",
    "election", # Queries that discuss the validity
       of the results of specific U.S. elections
    "immigration",
    "music", # Queries about music artists, songs,
       albums, concerts, etc.
    "other",
    "television", # Queries about movies, TV shows,
       films, plays, etc.
    "sports", # Queries about national sports and
       famous athletes
    "vaccine" # Queries about diseases, Covid,
       measles, vaccinations, Fauci, other public
       health issues pertaining to disease control,
       etc.
]
### Output Format
Output labels and very concise reasoning in a json
   format. Fill in the placeholders below:
{{"topics":["[tag 1]","[tag
   2]",...],"reasoning":"[reasoning]"}}
Do not add any additional characters or whitespace.
```

A.2 Elastic net classifier

Recall that $D_i = \{(e_j, y_j)\}_{j=1}^{n_j}$ denotes the labeled data for topic area i, where $e_j \in \mathbb{R}^{4096}$ is the embedding representation of claim j and $y_j \in \{0, 1\}$ indicates whether the claim is false.

The elastic net is a regularized logistic regression that combines ℓ_1 (LASSO) and

ℓ_2 (ridge) penalties. I estimate

$$\hat{\beta}_i^{EN} = \arg\min_{\beta} \left\{ \sum_{j=1}^{n_j} \mathcal{L}(y_j, e_j^T \beta) + \lambda_i \left[\alpha_i ||\beta||_1 + (1 - \alpha_i) ||\beta||_2^2 \right] \right\} \tag{25}$$

where $\mathcal{L}(\cdot)$ denotes the binary cross-entropy loss. The ℓ_1 penalty shrinks irrelevant embedding dimensions to zero, and the ℓ_2 penalty stabilizes estimation when predictors are correlated. The hyperparameters λ_i (overall penalty strength) and α_i (mixing parameter) are selected jointly via five-fold cross-validation to minimize classification error. The predicted probability that tweet t contains misinformation is given by:

$$\hat{p}_{it}^{EN} = \sigma(e_t^T \hat{\beta}_i^{EN}) \tag{26}$$

$$= \frac{1}{1 + \exp(-e_t^T \hat{\beta}_i^{EN})} \tag{27}$$

A.3 Random forest classifier

The random forest is an ensemble of B decision trees, each trained on a bootstrap sample of the data. For bootstrap iteration b, let $D_i^{*(b)}$ denote a sample of n_i observations drawn with replacement from D_i. A decision tree T_b is grown on $D_i^{*(b)}$, under the condition that at each split, only a random subset of $m < 4096$ embedding dimensions can be considered. The number of candidate features at each split is $m \in \left\{ \lceil \sqrt{p} \rceil, \lceil \log_2 p \rceil \right\}$. Splits are chosen to maximize information gain and reduce entropy. Suppose a node S contains $|S|$ tweets, with share p false. Then the entropy of S is given by:

$$H(p) = -p \log(p) - (1 - p) \log(1 - p) \tag{28}$$

and the selected split maximizes the entropy decrease:

$$\Delta H = H(S) - \left(\frac{|S_L|}{|S|} H(S_L) + \frac{|S_R|}{|S|} H(S_R) \right) \tag{29}$$

where S_L and S_R are the two new nodes created by the split. The predicted probability that tweet t contains misinformation is given by the proportion of trees that classify tweet t into the "false" class:

$$\hat{p}_{it}^{RF} = \frac{1}{B} \sum_{b=1}^{B} \hat{p}_b(y = 1 \mid e_t),$$

where $\hat{p}_b(y = 1 \mid e_t)$ denotes the fraction of training observations labeled "false" in the terminal node of tree b containing e_t. Hyperparameters — including the number of trees, the maximum depth of the tree, and the number of features considered at each split — are selected by maximizing out-of-bag accuracy over a predefined grid of candidate values.

A.4 Distribution of predicted probabilities

Figure 16 plots the distribution of the predicted probability that a given tweet in the Geotweet Archive is true. To obtain these probabilities, I select the model associated with each topic area that had the highest AUC, and then obtain either $\hat{p}_{it}^{EN}$ or $\hat{p}_{it}^{RF}$ by conducting inference using the corresponding model. Notice that the distribution for the predicted probabilities for vaccine-related tweets is substantially more concentrated near 0 and 1 than the corresponding distributions for the other topic areas. This pattern indicates stronger class separation in the embedding space for vaccine-related content, providing suggestive evidence that the vaccine classifier may be better performing when applied to the Geotweet Archive than the climate classifier or election classifier.

B Construction of the newspaper coverage index

The coverage index is constructed using the following four-step procedure. Each subsequent step is employed only if the previous conditions are not satisfied.

(1) The coverage index equals the weekly publication frequency, if such information is provided for a given newspaper-month in my panel.

(2) The coverage index equals the most recent weekly publication frequency, if the weekly publication frequency is provided for a previous newspaper-month in my panel, and a coverage change has not been recorded in the intervening periods.

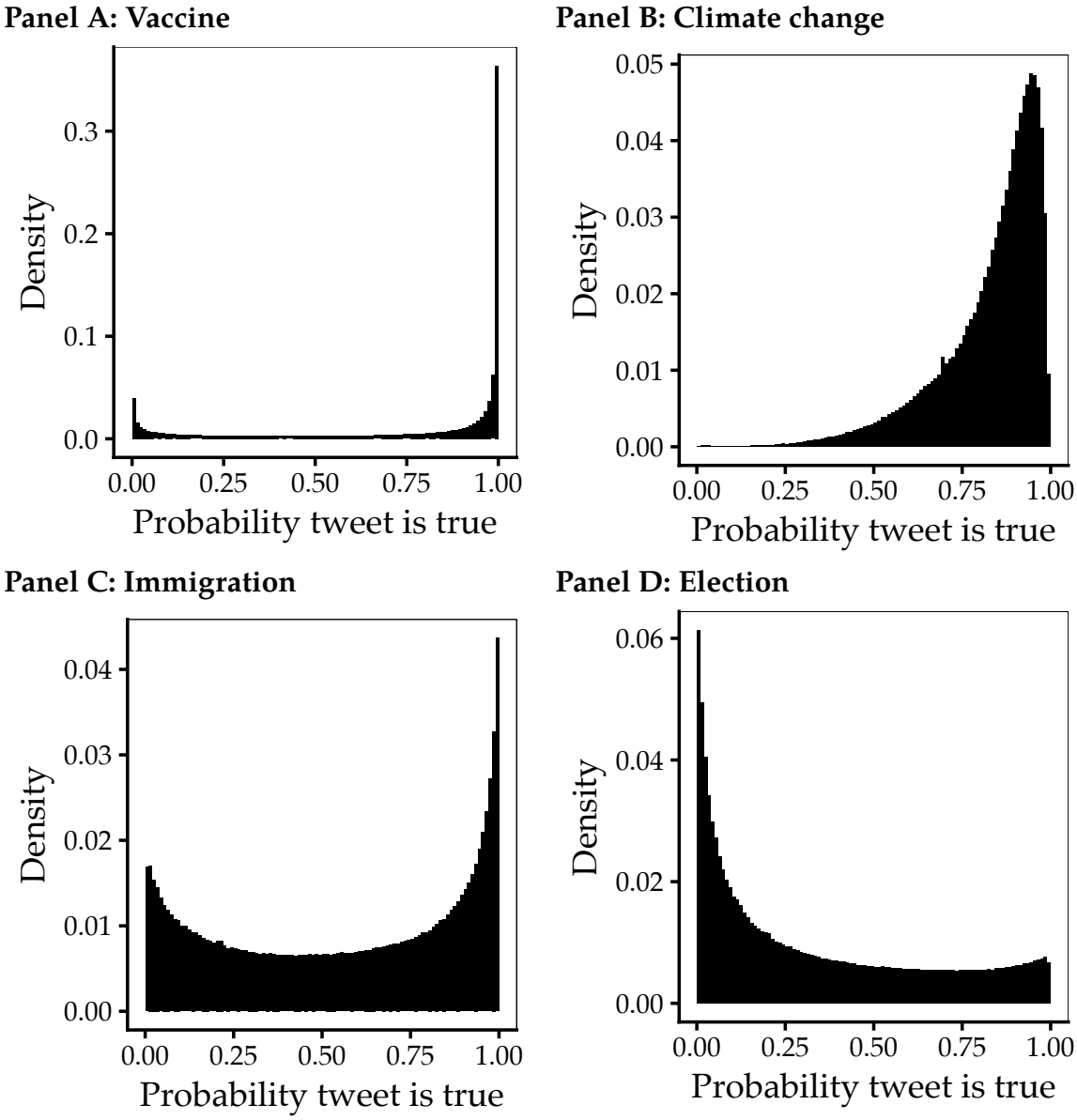
(3) The coverage index is predicted from the following set of variables: an indicator for whether the paper is a daily or a weekly, an indicator for whether the paper has an online presence, and the number of articles published on NewsLibrary.com (if available). I conduct this step using an elastic net model similar to the one defined in A.2. To train this model, I use the 291 hand-collected newspapers, for which I have (nearly) perfect information.

(4) If none of the above variables are present for a given newspaper-month, the coverage index equals 6 if the most recent observation for a given newspaper indicates that it is a daily, and the coverage index equals 2 if the most recent observation for a given newspaper indicates that it is a weekly.

C Treatment effects on likes and reposts

This appendix presents event study aggregates of ATTs on the total number of likes and reposts for the 24 months (2 years) before and after a county experiences a coverage change. I implement the methodology in section 5.

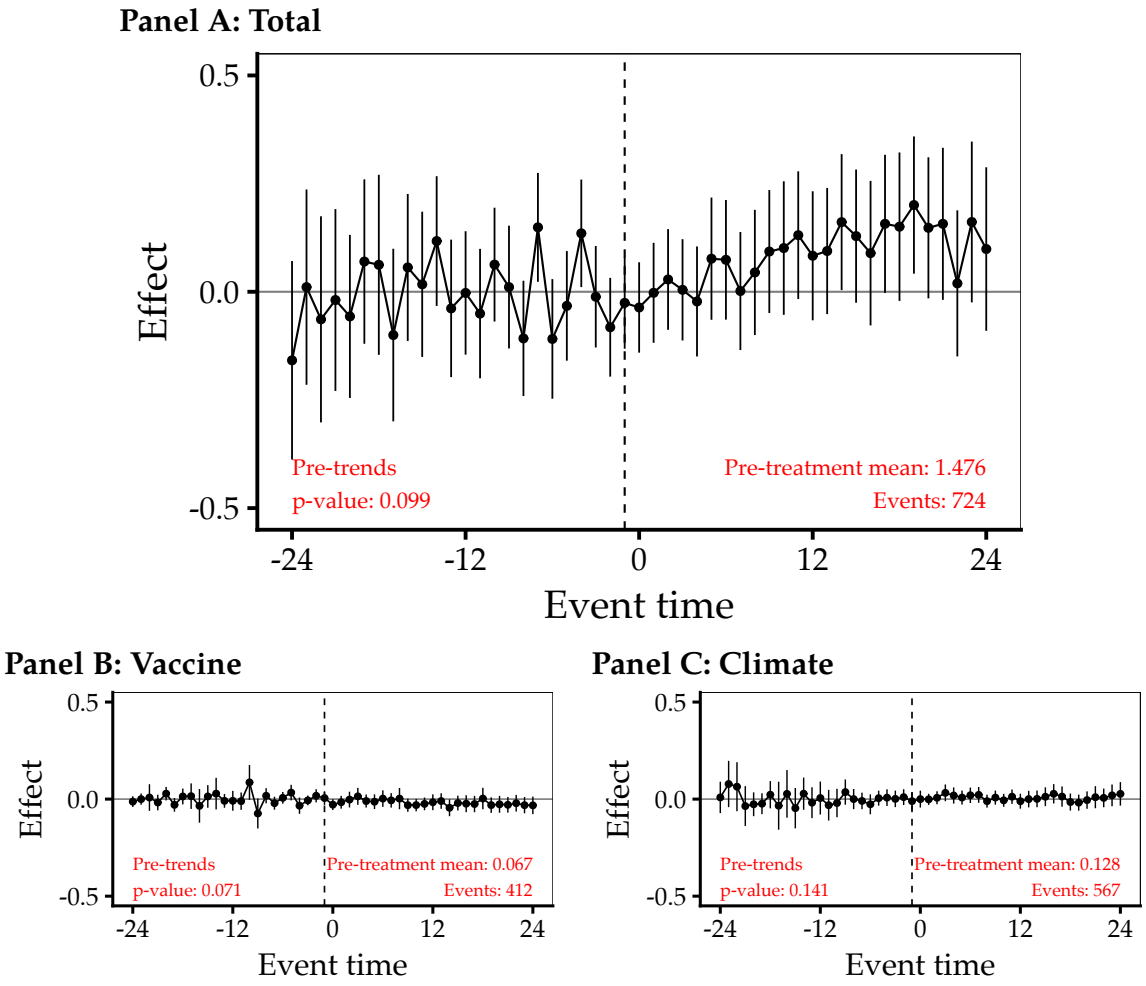
Figure 16: Distribution of predicted probabilities



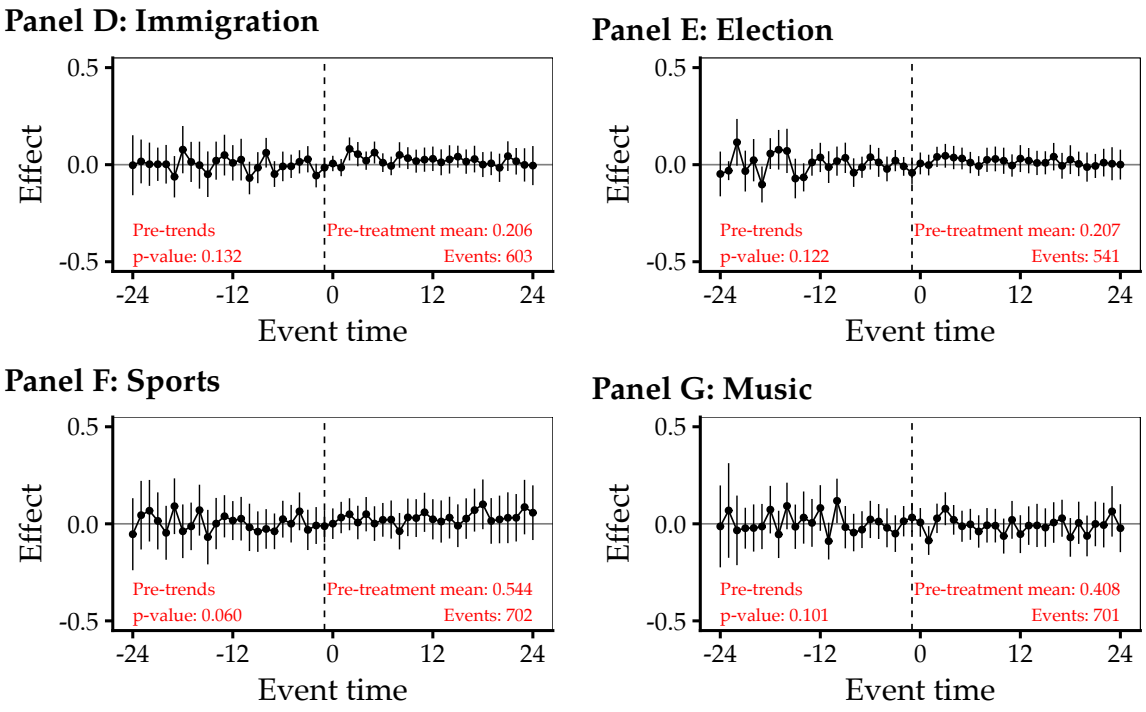
Notes: This figure plots the distribution of predicted probabilities that a tweet contains misinformation. These predicted probabilities are computed by conducting inference using the embedding of each tweet in the Geotweet Archive that pertains to a given topic. I apply the supervised classifier with the largest AUC on the holdout set for each topic area.

Figure 17 presents results for total likes. In panel A, the outcome variable, $Y_{i,t}$, is the modified log of the total number of likes for county i in period t. In panels B-E, the outcome variable is the modified log of the total number of likes on tweets about each of the four topic areas I have identified as prone to misinformation. In panels F and G, the outcome variable is the modified log of the total number of likes about each of the two placebo topic areas I have identified as likely to remain unaffected by changes in local newspaper coverage.

Figure 17: Event study aggregates of ATTs, total likes



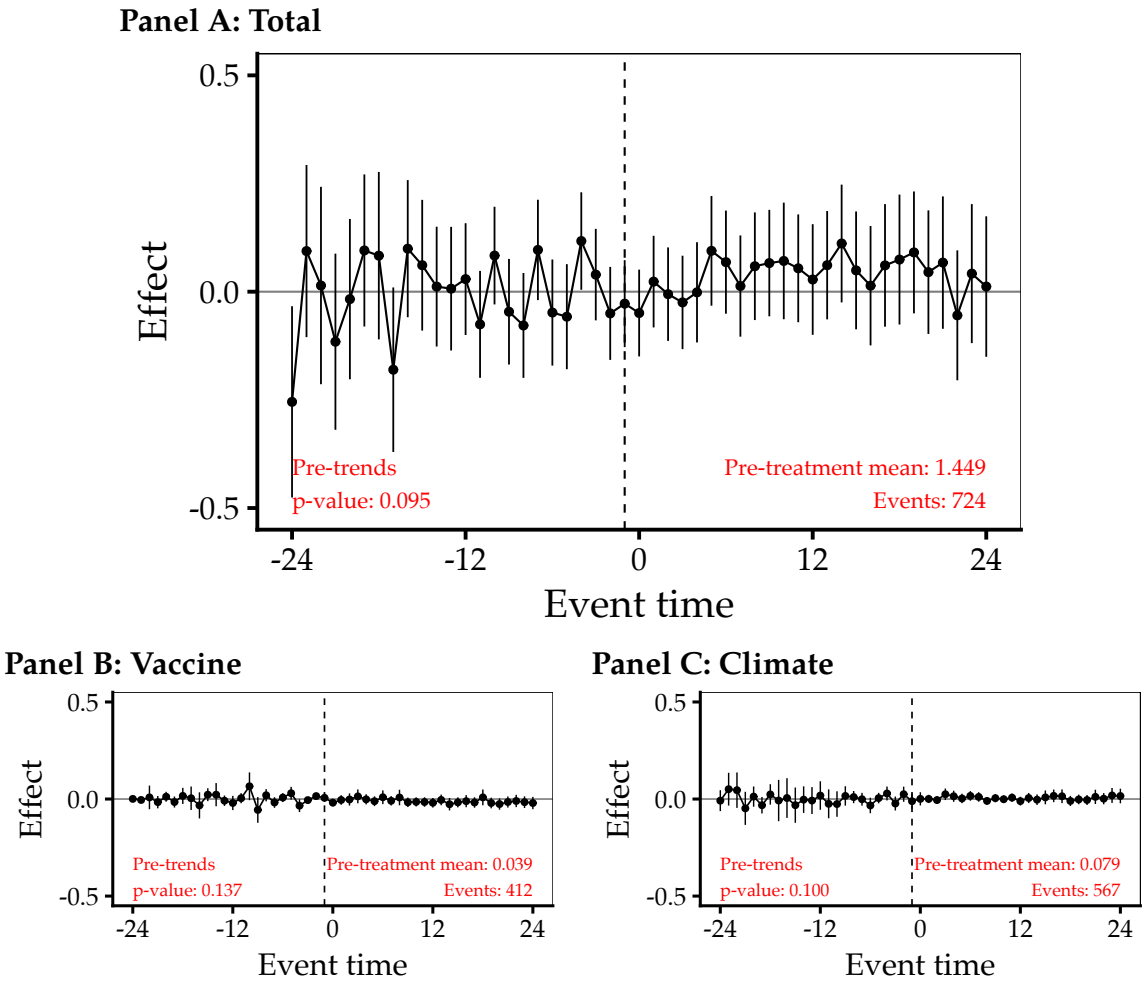
I do not find clear significant effects for any of the topic areas, though there does appear to be suggestive evidence that total likes across all topics may slightly increase. However, I am unwilling to conclude this definitively, given that I recover null effects for all topic areas of interest, and that the plots for the topic areas of interest do not look meaningfully different from the plots for the placebo topic areas. I fail to reject the null hypothesis of no pre-trends for all panels.



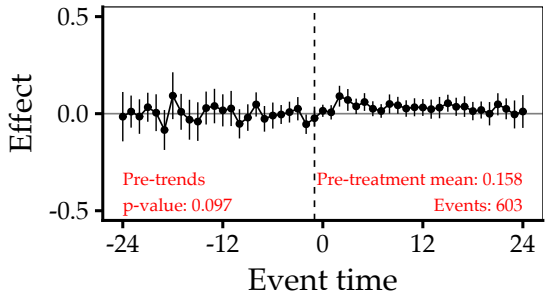
Notes: This figure plots "event study" aggregates of ATTs for the 24 months (2 years) before and after a county-level coverage change. The ATT parameter and aggregation procedure is described in detail in section 5 of the text. Panel A provides event study plots for overall Twitter activity. The dependent variable is the modified log of the total number of likes that tweets received in a given county in a given month. The modified log function is described in section 5 of the text. Panels B-E provide analogous event study plots, except the dependent variables only include tweets that pertain to topic areas I have identified as prone to misinformation. Panels F and G provide analogous event study plots, except the dependent variables only include tweets that pertain to the two placebo topic areas I have identified as likely to remain unaffected by changes in newspaper coverage. Annotations are provided in red to contextualize treatment effects for each figure. I pool the dependent variable for all ever-treated counties across all pre-treatment periods and report the mean. The number of "events" refers to the number of county-level changes in coverage that are used in estimation. The test for pre-trends is computed by conducting a Wald test using the bootstrapped covariance matrix. For each panel, I exclude from the sample all ever-treated and never-treated counties with fewer than 100 tweets in the relevant topic area over the entire sample period. Error bars represent pointwise 95 percent confidence intervals. Standard errors are clustered at the county level.

Figure 18 presents analogous estimates for total reposts. The outcome variable, $Y_{i,t}$, is the modified log of the total number of reposts for county i in period t. The conclusions are the same.

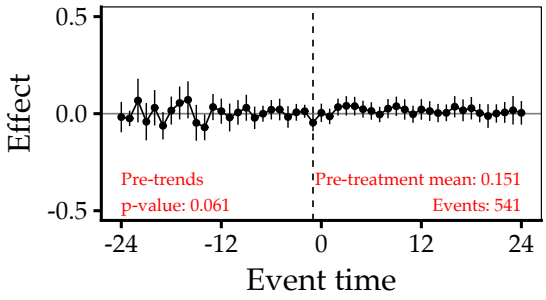
Figure 18: Event study aggregates of ATTs, total reposts



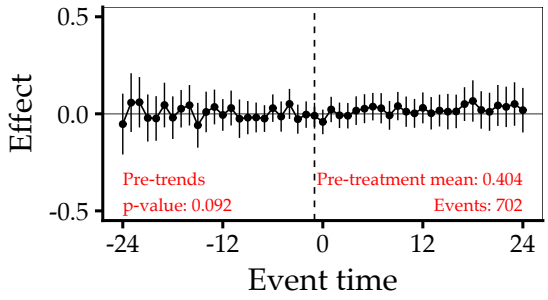
Panel D: Immigration



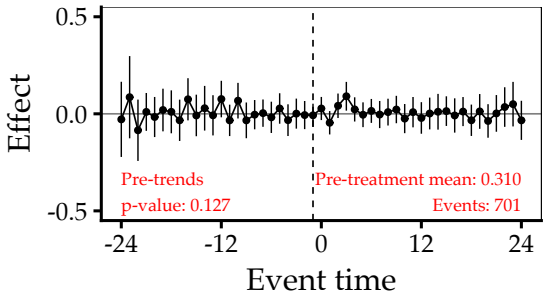
Panel E: Election



Panel F: Sports



Panel G: Music



Notes: This figure plots "event study" aggregates of ATTs for the 24 months (2 years) before and after a county-level coverage change. The ATT parameter and aggregation procedure is described in detail in section 5 of the text. Panel A provides event study plots for overall Twitter activity. The dependent variable is the modified log of the total number of reposts that tweets received in a given county in a given month. The modified log function is described in section 5 of the text. Panels B-E provide analogous event study plots, except the dependent variables only include tweets that pertain to topic areas I have identified as prone to misinformation. Panels F and G provide analogous event study plots, except the dependent variables only include tweets that pertain to the two placebo topic areas I have identified as likely to remain unaffected by changes in newspaper coverage. Annotations are provided in red to contextualize treatment effects for each figure. I pool the dependent variable for all ever-treated counties across all pre-treatment periods and report the mean. The number of "events" refers to the number of county-level changes in coverage that are used in estimation. The test for pre-trends is computed by conducting a Wald test using the bootstrapped covariance matrix. For each panel, I exclude from the sample all ever-treated and never-treated counties with fewer than 100 tweets in the relevant topic area over the entire sample period. Error bars represent pointwise 95 percent confidence intervals. Standard errors are clustered at the county level.