Measuring the Effect of Social Media Usage on Polarization

Freek van Sambeek*

May 5, 2023

Click me for the latest version

I propose a method of quantifying polarization using AI classification models which does not rely on predefining groups of affiliation. Using variation in social media use caused by social media outages, I find that social media usage does not have a significant causal effect on polarization, but that larger existing AI classification models would produce substantially more precise and reliable results. Lastly, I find that a polarization classifier trained on Reddit conversations does not accurately identify polarization in Twitter interactions.

1. Introduction

The idea that social media usage increases polarization is popular mainstream: Pew Research Center (2022) reports in a global representative study that about 65% of people think social media has made people divided in their political opinions, with only 8% of people thinking the opposite. In the US, the percentage of people thinking social media makes people politically divided was the highest, standing at 79%. Similar to the divide in people's political opinions is affective polarization: the extent to which people feel more negative toward other major political parties than their own (Iyengar et al., 2019). Affective polarization is undesirable because it can decrease the efficacy of government (Hetherington and Rudolph 2015). Knowing whether the truth aligns with the main public sentiment in this case is therefore important, potentially motivating a policy response.

In this paper, I measure the causal effect of social media availability on affective polarization in the United States, United Kingdom, Canada, and Australia. I start by creating a dataset of all conversations on Twitter linked to any of these countries from the start of 2018 to the end of 2022. I then classify the topic of discussion and the polarization score of the collected interactions using fine-tuned large

*London School of Economics and Political Science (email: f.van-sambeek@lse.ac.uk). I would like to thank Rachael Meager for her supervision, Matthew R. Levy for his guidance, Judith C. Shapiro for her mentoring, my fellow EME students for their support, and Kostas Kalogeroupoulos, Marcos E. Barreto, Xinghao Qiao for their comments.

language models. I aggregate the polarization scores to compile a dataset of day-level polarization by country and discussion topic, weighting interactions based on an algorithm that considers the likes received by each tweet of an interaction. I then identify the duration and extent of technological outages for any of the major social media platforms to identify a causal relationship between social media availability and affective polarization.

I find the causal effect of social media usage on political polarization to be insignificant. Nevertheless, the usage of bigger, more expensive classification models of OpenAI would identify variation in the outcome variable more accurately and give rise to more precise and reliable coefficient estimates. Lastly, I find that a neural network polarization classifier trained on a large dataset of labelled vector-embedded Reddit conversations from a selection of discussion forums is not accurate in predicting polarization from vector-embedded Twitter conversations.

This paper contributes to the literature evaluating the effects of social media usage on polarization. Previous research to investigate a potential effect of social media usage on affective polarization employs field experiments[1], event studies[2] or online conversations from a specific type of users[3]. In all studies like these, the sample is limited in either (i) the size[4] or the representativeness of the group of individuals in the data, (ii) the range of topics in which polarization is measured, or (iii) the time period to which the analysis is relevant. These limitations lead to overly local results based on non-representative data, which means the estimated effects may be biased relative to those that the authors are actually interested in finding. This paper instead evaluates the polarizing nature of social media by studying of interactions on a large scale over multiple years, across an exhaustive list of discussion topics, and among different countries.

This paper also contributes to the literature measuring polarization. The method I use to measure polarization does not make an effort to infer to which one of two groups any party in the interaction belongs. I choose this approach because, polarization is multi-dimensional in the sense that there may be different distinct groups or affiliation per topic, and perhaps more than two groups of distinct affiliation for some topics (McCoy et al. 2018). The popular approach of calculating polarization after identifying two groups of affiliation in the data does not account for this complexity and may therefore be too simplistic to accurately display the variation in polarization that policymakers care about.

This paper has several important limitations. First, the causal effects I estimate should be interpreted as the effect of short-lived reductions in social media availability. This is because widespread social media outages have never lasted more than a day. Perhaps the effect of social media availability is significantly non-linear in the hours it is made unavailable to the public for use. Second, the outcome

[1]For example, Ro'ee (2021) shows experimentally that random variation in exposure to news on social media substantially affects the slant of news sites that individuals visit.

[2]Take Lee et al. (2018), who attempt to show that social media indirectly contributed to polarization through increased political engagement by analysing changes in political views using panel data collected in South Korea between 2012 and 2016, a time of rapidly increasing social media adoption.

[3]E.g. Barbera (2014), who limits their analysis to users who follow political Twitter accounts.

[4]This is problematic as measuring the treatment effect to a randomly, representative sample assumes the non-treated individuals in the overall population do not affect the overall effect while people in the sample interact with non-treated individuals in their daily lives.

variable of affective polarization is constructed from interactions from Twitter only, which will limit the external validity of the causal estimates. Nevertheless, I argue that the algorithm I use to aggregate interaction-level polarization scores into a panel dataset of polarization produces a representative measure of polarization experienced among the Twitter population. With about one in four people actively using Twitter in the US, UK, Canada, and Australia in 2023 (Pew Research Center 2023; Kepios 2023), the results of this paper would still hold with respect to a large chunk of the populations in the countries of interest.

The remaining sections 2–7, respectively, present the theoretical framework, data, empirical strategy, results, discussion, and conclusion.

2. Theoretical Framework

I begin with a qualitative theoretical framework to fix ideas about the relationship between social media use and polarization in a society. I interpret affective polarization to be to what extent the groups in an average interaction in a society are not willing to listen to each other and acknowledge the arguments and opinions of each other, because this corresponds to the average antagonism an individual in a society experiences and is easy to interpret. The hypothesis is then as follows. Assume first that social media algorithms sort users into echo chambers, with users becoming less likely to acknowledge and listen to the opinions and arguments of randomly chosen individuals in the same society the more they use social media. If this effect materializes in day-to-day interactions between the citizens of a society, the consumption of social media by one person can make others more antagonistic, too, by spillover effects: one can expect an individual's marginal consumption of social media to have an effect on their attitudes towards people around them, which diffuses from that individual over all interactively connected individuals, manifesting in a change in polarization in the given society.

Now, whenever a social media service goes down, like Facebook, there are some people that were going to use Facebook during this outage who suddenly can not use Facebook anymore. They will spend their time in some other way for the duration that they were going to use Facebook. A decent share of the people that were going to use Facebook during the outage will probably switch to another media type to consume which is available to use. It is safe to say that not all of these people will substitute their Facebook usage fully with the use of another social media service. The non-social media substitutes will therefore be consumed to some extent. Suppose each of these substitutes is more or less polarizing than Facebook.

Aggregating over all substitutes using a weight proportional to the share of people that switch to each one, we can speak of Facebook substitutes being more or less polarizing than Facebook. Policymakers are probably keen to see whether the average Facebook substitute is more or less polarizing than Facebook, as this would allow them to conclude whether and how the availability of a social media service contributes to the levels of polarization that people experience in their societies. On Twitter, changes in the share of interactions that are polarized should consequently be visible. I try to capture just this change in the nature of interactions on Twitter for five of the major social media

services in English-speaking countries: Facebook, Instagram, YouTube, TikTok, and Snapchat. I do not study the effect of Twitter usage on polarization because no tweets can be posted during a Twitter outage.

3. Data

3.1. Tweets

I opted to collect data from Twitter as it is a platform of publicly available interactions that about one in four people in the US use, and will be similar in the UK, Canada, and Australia based on population penetration rates (Pew Research Center 2023; Kepios 2023). This is important as I aim to study the effect of social media use on polarization at a larger scale than other studies do. I only collect tweets from the US, Canada, the UK, and Australia, as a large corpus of tweets can only be feasibly analyzed computationally and models for computational analysis of language tend to perform best on English text (Muennighoff et al. 2022).[5]

I focus on tweets posted in the last five calendar years, 2018 to 2022, because the social media recommendation systems and content type have probably changed much since before 2018 and policymakers would care about the current state of social media's polarization. I filter Tweets by country of origin because outages can sometimes be location-specific (FirstPost 2023). This limits the sample, though, as tweets are not automatically geo-tagged, meaning the only tweets in the data used are those from users who manually turned on geo-tagging of their tweets, which constitutes about 2% of tweets (Twitter 2021). This reduces the representativeness of the results but prevents attenuation bias that would arise if one did not link outages to conversations from the place they occur in.

For all organic[6] tweets on Twitter for the given time range and geographic filters, I add a direct reply of another user to that organic tweet to create tweet-reply pairs. I then label the tweet-reply pair interaction to be polarized, neutral, or anti-polarized. I define such pairs to be neutral when not both the tweet and the reply are subjective (i.e. containing an opinion of sorts) or when there is disagreement, but this disagreement is constructive. I define a tweet-reply pair to be anti-polarized when there is clear/enthusiastic/warm agreement between the tweet and the reply. Lastly, I define a tweet-reply to be polarized when there is non-constructive disagreement or if one user displays a toxic sentiment towards the other user.

Identifying polarization on the scale of individual interactions is agnostic about the groups of affiliation that exist among the users while simultaneously allowing for any user to belong to any group. For each interaction between two people, both people are part of a group representing a contextual sentiment. If these groups are not composed of multiple people (by the likes the different users receive), they are composed of the two users themselves. This interpretation is motivated by the outcome variable of interest being to what extent the groups in an average interaction in a society are not willing

[5] For a more detailed description of my tweet collection methodology, see page 29 of the appendix.

[6] An organic tweet is either a standalone tweet or a tweet quoting another tweet.

to listen to each other and acknowledge the arguments and opinions of each other. In the literature, this would mostly correspond to trying to measure affective polarization[7] (Kubin and Von Sikorski 2021).

3.2. Identifying Polarized Interactions

I use[8] a language model to classify interactions for me, as it would not be feasible to manually label all collected interactions. I chose the two mainstream types of transformer-based AI models, which generally outperform any other type of model in classification tasks (Neelakantan et al. 2022).

1) The first method I use is embedding the text of tweet-reply pairs into a high-dimensional vector with a pre-trained embedding model. These vector embeddings will have meaning in the sense that similar pieces of embedded text represent points relatively close to each other in the high-dimensional vector space. Complex areas contained in the embedding space, like areas that identify whether the corresponding piece of text features a polarized interaction, can be unraveled[9] by fitting a flexible neural network to a training dataset of manually labelled tweet-reply pairs (the outcome variable) and the embedding vector of the corresponding text (the input variables). This neural network can then be used to classify the embedding vectors of tweet-reply pairs outside of its training dataset.

2) The second model I use is a large pre-trained[10] text-completion model GPT-3. I show this model a lot of examples of how a human classified the interaction type of certain tweet-reply pairs. After doing that, you can supply the model with new tweet-reply pairs to classify on its own. Given what it has seen, it will predict the outcomes of the new interactions.

As both of these methods require a dataset, I manually labelled around 1,000 tweet-reply pairs to be part of any of the following three classes: `anti-polarized`, `neutral`, and `polarized` according to the definitions listed in the previous section.

Some tweets in the sample do not have direct replies from other users, which I label to be neutral in their interaction type. I decide not to drop these tweets because they contain information about people's interactions too: at least some of these tweets are shown on other users' feeds[11]. I label these tweets as `neutral` as no user that saw the tweet felt a strong enough need to reply[12], meaning their

[7] Affective polarization is equivalent to out-of-group animosity, the "toxic" component of polarization, whereas ideological polarization is related to the extent of between-group differences which does not directly imply animosity (Kubin and Von Sikorski 2021).

[8] For more information on the reasons I chose these models over existing methods in political economy research, see page 29 of the appendix.

[9] As long as the model that creates these vector embeddings is 'good enough' in the sense that it creates these embeddings with this information contained in them.

[10] Pre-trained means that a model has been fitted to data already. This pre-training may be transferrable across different use cases, especially for large language models like GPT-3 that have been trained on a substantial share of the internet and books in the world.

[11] A feed is the home page on which Twitter displays a selection of tweets consisting of tweets of people that you follow and tweets it recommends to you.

[12] See the discussion for why the existence of deleted tweets does not pose a significant threat to this argument.

attitude towards the tweet probably was not extreme to either end of the spectrum used here, polarized or anti-polarized.

Text-completion models continue text in the way they estimate to be most likely, so it is a best practice to give the model a prompt[13] that is most reminiscent of what it has seen on the internet and in the books that it has been trained on (OpenAI 2022). The tweet-reply pairs should therefore be formatted in a way such that the model is most likely to continue the text of a tweet-reply pair with the correct label of the labels you want the model to classify interactions with.

Given a tweet with the text `[tweet]` and a reply with the text `[reply]`, I therefore give the following prompt to the text-completion model for each tweet-reply pair:

> *"Person 1: '`[tweet]`' Person 2: '`[reply]`'*
>
> *Interaction type between Person 1 and Person 2: "*

This makes it clear that `[tweet]` and `[reply]` are utterances that come from two different people, emphasizes that what follows after the prompt is a word for the type of interaction between these two different people, and includes whitespace to avoid concatenating a word at the end of the tweet with a word at the beginning of the reply.

The training dataset has around a thousand of such examples labelled by one of `disagreement`, `neutral`, and `agreement` that correspond to `polarized`, `neutral`, and `anti-polarized`, respectively. I chose this mapping because the usage of the words "disagreement", "neutral", and "agreement" is probably more prominent in the internet and literary corpora that GPT-3 was trained on. This means that completing the prompt as shown above correctly will be more "natural" to the model and limit the noisiness of inputs to the model. The thousand or so examples should then serve to convey that constructive disagreement should be classified as `neutral` and toxic interactions should be classified as `disagreement`, which are distinctions that need to be made in order for the classes `disagreement`, `neutral`, and `agreement` to correspond to `polarized`, `neutral`, and `anti-polarized`, respectively.

I neutralized links and usernames to prevent the text-completion model to make inferences on the interaction type based on words that just so happen to be in any of the links or usernames.

For information, if `[tweet]` is a tweet that quotes another tweet with the text `[quoted]`, I change the prompt to:

> *"In response to: '`[quoted]`' Person 1: '`[tweet]`' Person 2: '`[reply]`'*
>
> *Interaction type between Person 1 and Person 2: "*

Because the quoted tweet sometimes provides context necessary to determine the interaction type between the tweet and the reply.

[13] A prompt is the text you give to the pre-trained text-completion model to continue.

3.3. Aggregating Polarization Scores

After formatting all tweet-reply pairs and labelling the interaction type for each with the fine-tuned text-completion model, I convert each label to a numeric label in which `polarized`, `neutral`, and `anti-polarized` correspond to the classes `1`, `0`, and `-1`, respectively. This allows me to take a weighted average over all interactions to distill the interactions in a given country on a given day into one polarization score. I aggregate the interactions using weights because social media downtime data is daily and heterogeneity in the number of likes between tweets conveys information about the number of people involved in an interaction.

I weigh interactions as follows, given the number of likes of the tweet and the number of likes of the reply:

Table 1: POLARIZATION AGGREGATION WEIGHT BY INTERACTION TYPE

	Anti-polarized	Neutral	Polarized
Pairs	W	W	$W \cdot \frac{min(likes_{r,w},\ likes_t)}{max(likes_{r,w},\ likes_t)}$
Single	Not applicable	$likes_t$	Not applicable

Notes: This table presents the weights I use to aggregate the polarization scores of interactions into country × date × topic-level polarization estimates. The cells that show "Not applicable" do so because single tweets can not be involved in an anti-polarized or polarized interaction, and do therefore not exist.

Where:

- Cells in the row of "Pairs" correspond to the weight of the numeric polarization label for a tweet-reply pair given the class of the interaction type, whereas cells in the row of "Single" corresponds to the scaling of the numeric polarization label for a tweet with no direct replies from a different user.
- $likes_t$ corresponds to the like of the tweet, whereas $likes_r$ corresponds to the like of the reply.
- $\mu(x)$ takes the empirical mean of x.
- $likes_{r,w} := likes_r \cdot (\mu(likes_{t,p})/\mu(likes_r))$.
- $likes_{t,p}$ is the number of likes of a tweet that is part of a tweet-reply pair.
- $W := likes_{r,w} + likes_t$.

$likes_{r,w}$ is the number of likes of a reply, weighted to make it comparable to the likes of the tweet obtained by scaling $likes_r$ by the ratio between the average number of likes of the non-reply tweet in a tweet-reply pair and the average number of likes of a reply in a tweet-reply pair. So this scaled version of $likes_r$, $likes_{r,w}$, is comparable to $likes_t$, as they have the same averages.

The intuition behind W is that it represents the number of people involved in an interaction. Namely, as a user that has seen a reply to a tweet is more likely to have seen the parenting tweet than a user that

has seen the parenting tweet has seen that particular reply, treating likes on a reply equal to likes on a tweet would be unfair. Simply adding $likes_{r,w}$ and $likes_r$ together is motivated by the assumption that on average users have a similar likelihood of liking a tweet relative to the likelihood of liking a reply when supplied with any tweet-reply pair, and that users who like a reply feel a similar kind of need to endorse to users who like a tweet.

I also multiply W by $\frac{min(likes_{r,w},\ likes_t)}{max(likes_{r,w},\ likes_t)}$ for polarized interactions: the ratio of the minimum of the comparable figures of $likes_t$ and the scaled version of $likes_r$ to the maximum of the two. This way, the more an interaction is divided, i.e. the more comparable the endorsement received by each side of an interaction, the higher this fraction will be. A polarized interaction with relatively few endorsing one user in the interaction means this fraction is smaller, as the minimum term will be small relative to the maximum term.

If all tweet-reply pairs in a day are maximally polarized, i.e. $likes_t$ and the weighted version of $likes_r$ are the same so that the minimum divided by the maximum equals 1 for all tweet-reply pairs, then the aggregated polarization score will be 1. If all tweet-reply pairs in a day are anti-polarized (and there are no tweets without any replies on that day–"single" tweets), the aggregated polarization score will be -1.

3.4. Classifying Interaction Topic

I classify tweets to be of any of the categories (i) `politics`, (ii) `entertainment`, and (iii) `advertisement`. These are defined as follows:

(i) `advertisement` is used as a label whenever the main tweet is spam or an advertisement of some sort.

(ii) The category `politics` is defined to contain any tweet in which anything is discussed that is relevant to the political state of the world, like financial or economic news, scandals of politicians, policy, political debates, and conversations about religion, norms and values, and philosophy in general.

(iii) `entertainment` contains all of the remaining. In the manual classification of categories for these topics I found that most tweets in this remaining class are likely to be about personal life, sports, and TV series.

I can then distinguish between a polarized interaction of two people about Brexit and one about a cat video. Policymakers probably care about that distinction. Furthermore, these categories allow me (a) to filter out all tweets from bots[14] and adverts to be able to focus on genuine human interactions, and (b) to choose to filter on interactions with a political dimension which would seem to feature most of the polarization of interest in contemporary societies, given the definition of "politics" used.

[14] Bots on Twitter are computer programs that algorithmically generate and publish tweets. It is estimated that the share of tweets tweeted by bots as a proportion of all tweeted tweets is around 21% (SimilarWeb 2023).

With these classes determined, I choose to fine-tune a large pre-trained deep-learning language model to classify my data, for the same reasons that I mentioned for the classification of disagreement. In this case, the following prompt template is applied to all records:

"Person 1: '`[tweet]`*' Person 2: '*`[reply]`*'*

Category of the above tweet(s): "

In case a record is classified without a direct reply from a distinct user, I use the following template:

"Person 1: '`[tweet]`*'*

Category of the above tweet(s): "

And again, in case the starting tweet quotes another tweet, I prepend the prompt by "*In response to: '[quoted]*' ".

3.5. Social Media Outages

I use social media outages to quantify the causal effect of changing usage of social media services on polarization. The reason for this is that technical issues are at the root of social media outages and that those technical issues can happen at any moment at the expense of any individual mistake, meaning they are volatile in their occurrence. For illustration, below is a list of possible causes of social media outages[15]:

1) Changes to the platform's internet infrastructure.
2) Software bugs.
3) Configuration changes to the platform's servers.

For the most part, outages are caused by updates to the servers of social media services. This is when teams of software engineers push changes to the algorithms that run behind the scenes to make a social media platform work. Namely, the bulk of what makes a social media platform run, recommendation systems, databases, notifications, and account functionality to name a few, is processed on computers different from users' ones. Whenever an update to any of these systems is carried out, there can be unforeseen implications to the overall workings of the platform. For some of these, the platform may be unavailable to a proportion of users[16]. As these systems are complex, it may not always be possible to ensure beforehand that certain updates are safe to implement on the overall scale of the platform, by which outages may occur.

One concern with using outages as exogenous variation in the usage of social media is that, with teams of software engineers being aware of the threat of an outage, these updates are carried out when

[15] For example, see Wikipedia (2023) for a detailed list of outages of Google services and their causes.
[16] Some outages are exclusive to certain regions or users.

the cost of an outage is lower, say during periods of lower usage of the platform. However, social media usage can be expected to be relatively stable over time, and perhaps even growing. Additionally, with large scale adoption of these social media services across many time zones, for an outage at any point during the day, there will be a large group of users affected, meaning waiting for the clock to hit 3 am US Eastern Standard Time before pushing server updates is not necessarily a sensible strategy for software engineering teams.

I use Google Trends data to identify outages. Google Trends is an online tool created by Google which allows users to view the popularity of any search query on Google over time. This is helpful for the analysis of interest for the following reason. According to Google's answers to Google Trends FAQs (Google 2023), for a given search prompt, a given time period, and a given geographical region, like a country, Google Trends returns a graph displaying the popularity of that search prompt over time in that country, where the popularity is indicated by the proportion of queries like the search prompt of all Google queries made in that country. This proportion is scaled such that the point of highest popularity in the graph is equal to 100. Assuming that the overall browsing volume in the countries of interest in this study is relatively stable over time, being able to see this proportion of search queries changing over time is informative because the number of people hit by a given outage should be projected nicely onto the related Google Trends chart (assuming people google for outages to a similar degree over time for an outage of similar severity).

I scrape[17] Google Trends data for:

(1) Each of the services: Facebook, Instagram, YouTube, TikTok, and Snapchat.

(2) Each of the largest mainly English-speaking countries: United States, United Kingdom, Canada, and Australia.

(3) The beginning of 2018 to the end of 2022.

I also collect data on Twitter outages for (1) and (2) in order to investigate whether Twitter outages sometimes overlap with outages of the platforms of interest.

The search query I use in this study is the query "*is [platform] down*", where [platform] corresponds to the name of the social media platform of interest. The reason for this search query is that by trial and error, it seems to be used most out of many alternative queries, and Google autocomplete suggests it as one of the top options when typing "*is [platform]*" in search engines when tested in freshly installed browsers (meaning there is no history stored by the browser of a researcher typing that into their Google search bar before). Capitalizing letters in a search query has no effect on the data returned by Google Trends.

I then fit the popularity data to the worst complete service outage for each service so that the maximum popularity of a search term for a given platform should correspond to the duration of the service's worst outage.

[17] For more information on the technical decisions I made when scraping search popularity data from Google Trends, see page 30 of the appendix. For downtime estimation alternatives, refer to the same page.

4. Empirical Strategy

4.1. OLS

I set up two regression models to measure the causal effect of social media availability on affective polarization. I use robust standard errors to account for positive autocorrelation in the error term which may be expected because of polarization levels depending on polarization in previous periods and because of possible time-trend-related heteroskedasticity. I do not include polarization levels from previous periods as regressors in this model because these do not cause outages and thus are not confounders. The specifications are

$$\text{polarization}_{c,t} = \text{effective\_outages}_{c,t}\beta + \Delta\text{twitter\_usage}_{c,t}\gamma + \kappa_c + \eta_t + \varepsilon_{c,t} \tag{1}$$

and

$$\text{polarization}_{c,t} = vec\{\text{Lagged\_Outages}_{c,t}\}'\beta + \Delta\text{twitter\_usage}_{c,t}\gamma + \kappa_c + \eta_t + \varepsilon_{c,t}, \tag{2}$$

where $\text{polarization}_{c,t}$ is a scalar variable that takes the outcome variable of the polarization level in country c at time t, where time is at the date-level. κ_c and η_t are country and year fixed effects. $\text{effective\_outages}_{c,t}$ is a row vector in which each element corresponds to the effective outage severity of a social media service at time t in country c. $\text{Lagged\_Outages}_{c,t}$ is a matrix containing the estimated outage for all services (columns) over the current and previous six days (rows). For a single platform with index i, the effective outage severity $\text{effective\_outage}_{i,c,t}$ is calculated as follows:

$$\begin{aligned}\text{effective\_outage}_{i,c,t} =& 1.2\left(\frac{1}{3}\text{outage}_{i,c,t} + \sum_{n=1}^{4}\left(\frac{1}{3}\right)^n \text{outage}_{i,c,t-n}\right)\\ \approx& 0.40\cdot\text{outage}_{i,t} + 0.40\ \text{outage}_{i,c,t-1} + 0.13\ \text{outage}_{i,c,t-2} + 0.05\cdot\text{outage}_{i,c,t-3}\\ &+ 0.02\cdot\text{outage}_{i,c,t-4}\end{aligned} \tag{3}$$

I choose this specification for interpretability reasons. Namely, (i) this variable can be seen as a weighted average of the closest lagged variables of outage severity[18], and (ii) the terms decay exponentially in their weight while the closest two lags are of equal weight given that outages occur during a day at which polarization is measured and the diffusive effect that follows from a shock to social media usage may take a bit of time to set in. Namely, the hypothesis discussed in section 2 is that the marginal availability of social media for an individual has some aggregate effect on polarization that diffuses through society from the people affected. Therefore, I model an outage to have an effect over a multiple of days after the shock, but decreasing in its effect over time and reaching zero after

[18] $1.2\left(\frac{1}{3} + \sum_{n=1}^{4}\left(\frac{1}{3}\right)^n\right) \approx 1.$

some set number of days. One may expect social media consumption to gradually revert to a level that is natural to the people in the society of interest, meaning that the polarization of interactions will gradually revert to equilibrium, holding fixed all variables after the end of an outage. All other variables can be seen to be held fixed in the regression of interest as I treat outages to be random in occurrence.

$\Delta\text{twitter\_usage}_{c,t}$ corresponds to the change in the popularity of the search term *"twitter.com"* on Google in country c at time t, relative to the popularity of that same query in the previous day, and inversely scaled by the popularity of Twitter over time (estimated by taking the yearly popularity of *"twitter"* over time from Google Trends in country c). This should adequately capture relative changes in the usage of Twitter, as a decent share of Twitter users uses Twitter through the website[19], meaning that the search intensity of *"twitter.com"* is appropriate to measure Twitter usage. I inversely scale the popularity of *"twitter.com"* by the popularity of *"twitter"* because it is the proportional change in Twitter usage that is a confounder, not the absolute change in Twitter usage.

I do not control for substitution to other social media services as it is hard to find accurate statistics on social media usage per service. The measured effect can therefore be interpreted as the effect of switching to the average substitute as a result of downtime of a particular social media platform, as discussed in section 2. If it is estimated that downtime of social media services reduces polarization and we assume that social media services are of similar polarizing nature to each other, then we can say that the average non-social media substitute is even less polarizing than the estimated treatment effect as the substitution to alternative social media usage is included in the estimated effect.

4.2. Bayesian Hierarchical Model

Additionally, I estimate the following Bayesian hierarchical regression model:

$$\text{polarization}_{c,t} = \text{effective\_outages}_{c,t}\beta_t + \Delta\text{twitter\_usage}_{c,t}\gamma + \kappa_c + \varepsilon_{c,t},$$

$$\varepsilon_{c,t} \sim N(0,\sigma),$$

$$\beta_t \sim N(\beta, \Sigma_{\beta_t}),$$

$$\beta \sim N(0, \Sigma_\beta),$$

$$\gamma \sim N(0, \sigma_\gamma),$$

$$\kappa \sim N(0, \sigma_\kappa),$$

$$\kappa_c \sim N(\kappa, \sigma_{\kappa_c}),$$

$$\Sigma_\beta \sim \text{diag}(\sigma_{\beta 1}, \sigma_{\beta 2}, \ldots, \sigma_{\beta 5}) + O,$$

$$\Sigma_{\beta_t} \sim \text{diag}(\sigma_{\beta_t 1}, \sigma_{\beta_t 2}, \ldots, \sigma_{\beta_t 5}) + O,$$

$$O_{i,j} = \begin{cases} \rho, & \text{if } i,j \in \{1,2\},\ i \neq j \\ 0, & \text{otherwise} \end{cases},$$

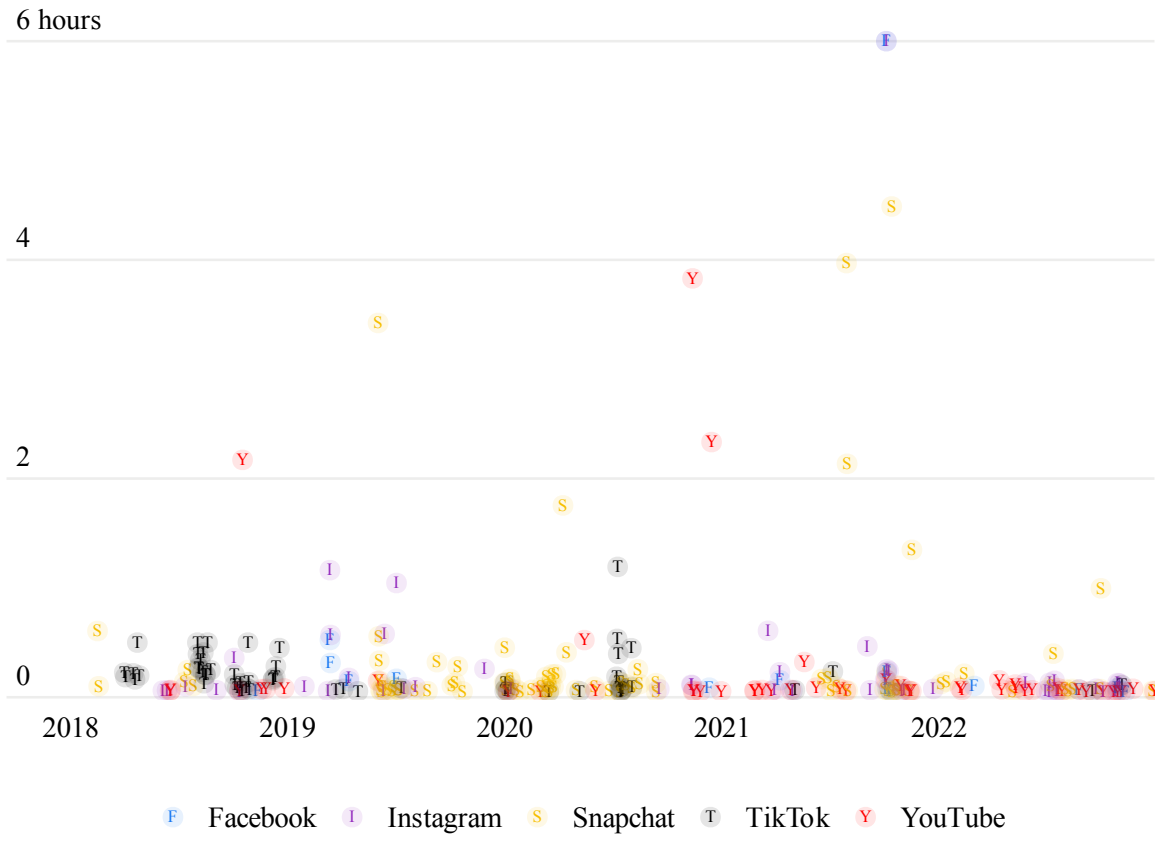
[19] Statista (2015) notes that about one-fifth of Twitter users use the platform through the web, which is has probably become less in the meantime, but should still mean variation in *"twitter.com"*'s popularity represents variation in usage for a large sample of Twitter users.

with all types of $\sigma \sim \text{Halfcauchy}(5)$.

β_t is a vector of coefficients per year, while κ_c is a country-fixed effect, from the hyperparameters β and κ, respectively. ρ is equal to the correlation between the effective outage severity of Facebook and Instagram, which are the first and second entries of the effective_outages$_t$ row vector. I use a diagonal covariance matrix for the error term because this makes the model estimation feasible and because model convergence is unlikely, otherwise.

Including ρ is motivated by the below picture of outages over time that displays little correlation between outages of any platform pair except Facebook and Instagram, which is additionally supported by Facebook and Instagram partially running on their parent company's servers (Meta's servers), which is not the case for any other platform pair.

Figure 1: Social Media Outages with Durations over the Years



Notes: This figure shows observations of the estimated outage durations across different social media platforms. The outage duration estimates are estimates of equivalent full platform outages. I include the observations from all countries. I use "equivalent" here to refer to the expected equivalent duration of a widespread outage (during which nobody in a country of interest is able to use the service) based on the popularity of the related outage search query on Google Trends. I do not include observations with an estimated equivalent outage duration of less than three minutes. Lastly, there were 9,130 observations (5 services with one observation over 5 years) before dropping all those for which the estimated equivalent outage duration was lower than three minutes, which left 254 observations for this graph. This means that there are about ten outages per service a year.

I fit the Bayesian model using Stan, which uses Markov Chain Monte Carlo algorithms to estimate

models. The priors on correlation matrices and variance parameters are chosen as recommended by the Stan manual (Gelman 2020).

The reason for estimating a Bayesian hierarchical model additionally to the OLS model is that country-fixed and time-fixed effects may deal with the country-specific and time-specific concerns, but do not use as much information as is available. Namely, one could argue that the average polarization in a country as is defined is determined by a lot of variables, but that a lot of these variables are similar for the English-speaking countries in the sample. Therefore, we may want to say that these country-specific base levels of polarization in the time frame used come from some not-too-disperse distribution. Additionally, as the concern related to time stems from social media platform algorithm updates over time, but these algorithm changes can be argued not to have been drastic enough to completely change the way these social media services work (in the time frame of 2018 to 2022), one may say that it is reasonable to model the coefficients to be drawn from a not-to-disperse distribution, too. This approach of assigning priors uses available information to impose structure while allowing for variation in the coefficients in the dimensions that potentially cause bias when fixed.

5. Results

5.1. Identification Polarized Interactions

I classified the interaction type of the tweet-reply pairs with a large pre-trained language model in two ways. First, with text embeddings and second, with text completion. The first method requires training a neural network on correctly labelled input data. As neural networks require a lot of training examples to do well on classification tasks with considerable variation in the input data (Brownlee 2019), achieving reasonable prediction accuracy would require manually labelling a substantial share of the tweet-reply pairs. Therefore, I use a publicly available dataset. I did not find any datasets containing polarization-labelled tweet-reply interactions, but what came the closest was a dataset called Debagreement (Pougu et al. 2021), which consists of Reddit post-reply pairs either labelled with `disagreement`, `neutral`, or `agreement`. These classes do not exactly correspond to the classes used in this study: `polarized`, `neutral`, and `anti-polarized`: going back to the definition of these classes in section 3.2, one can see that only some of the Reddit post-reply pairs that are classified as `disagreement` should belong to the class `neutral` before setting the former class to the latter. Debagreement consists of around 40,000 labelled interactions from five different forums on Reddit: r/BlackLivesMatter, r/Brexit, r/Climate, r/Democrats, and r/Republican. One can note that these interactions do not span the majority of interactions one may find on Twitter, maybe not even those about politics on Twitter. This means that a neural network trained on vector embeddings of interactions contained in this dataset may not do well out-domain on vector embeddings of interactions on Twitter.

I generate embeddings by leveraging OpenAI's state-of-the-art embeddings model (released December 2022), which I chose because it scores best out of other text embedding models when used on a varied collection of benchmark datasets for classification tasks (OpenAI 2022).

While the neural network[20] after training did okay on the validation data[21] from the Debagreement dataset, it did quite poorly on the embeddings of the tweet-reply pairs in the validation dataset of the data used in this study. The below confusion matrix shows the decomposition of the model's predictions conditional on any given true class.

Table 2: CONFUSION MATRIX EMBEDDINGS POLARIZATION CLASSIFIER

Actual \ Predicted	Polarized	Neutral	Anti-polarized
Polarized	8	4	6
Neutral	24	70	116
Anti-polarized	3	1	18

Notes: This table presents the raw Twitter interaction polarization classification accuracy for a neural network trained on a dataset of polarization-labeled vector embeddings of Reddit interactions. Each cell shows the total number of validation records belonging to the given categories. The red numbers represent incorrect predictions, the black numbers represent correct predictions.

From the above confusion matrix, one can tell that the classification accuracy is quite poor, about 40%. A degenerate model that always predicts `neutral` would do substantially better than this model in terms of the proportion of records it classifies to be correct. Therefore, using this embeddings-trained neural network would create significant noise in the outcome variable which would result in higher standard errors on the coefficients of the regressors of interest.

For the second method of classification of polarization, I use another model from OpenAI, the fine-tuning model Ada (OpenAI 2023). After I fine-tune this pre-trained language model on approximately 800 manually labelled tweet-reply pairs, which should be enough for this model to do well as I explain in section 3.2, the following is the confusion matrix of this model obtained by having it classify the roughly 250 records of labelled tweet-reply pairs set aside for validation of the model.

Table 3: CONFUSION MATRIX FINE-TUNED POLARIZATION CLASSIFIER

Actual \ Predicted	Polarized	Neutral	Anti-polarized
Polarized	3	11	4
Neutral	2	201	7
Anti-polarized	3	14	5

Notes: This table presents the raw Twitter interaction polarization classification accuracy of OpenAI's LLM Ada, fine-tuned to a training dataset of polarization-labeled Twitter interactions. Each cell shows the total number of validation records belonging to the given categories. The red numbers represent incorrect predictions, the black numbers represent correct predictions.

[20] I used a small and larger neural network on the training data. The former had one hidden layer of 800 neurons and the latter had three hidden layers of 2,500 neurons. All hidden layers had a dropout rate of 20% and used the ReLu activation function.

[21] Validation data is a random sample of the data that is not given to the prediction model at the training stage, so that after training one can estimate the accuracy of the model on data generated by the same process underlying the dataset.

Looking at the numbers on the diagonal in the above confusion matrix, the accuracy of this model is considerably better than the embedding classification model, about 80% versus 40%, respectively. However, what we may care about besides accuracy is whether variation in the outcome variable will be correctly identified. With the numeric mapping applied to the polarization labels, (i) given a `neutral` example one may want the model to estimate `0` on average, (ii) given a `polarized` example one may want the model to estimate a number strictly higher than `0` or given a `anti-polarized` example one may want the model to estimate a number strictly lower than `0`. In essence, one at minimum wants the average predictions given a true class to be ordinal, in the direction of the ordering of the classes as given by the numeric mapping. This is not the case for this model, as the average prediction of this model given a polarized interaction is lower than `0` instead of higher than `0`, and this average prediction is similar to the average prediction given an anti-polarized interaction. This is not a hopeful finding given the goal of correctly identifying variation in the outcome variable of polarization. However, because this validation sample is small, it is possible that the quality of these predictions is worse than they actually are for the entirety of the dataset.

The largest and most expensive fine-tuning text-completion model from OpenAI[22] outperforms the embeddings classification model and the smaller text completion model by a large margin, as can be seen in the confusion matrix below. It has an accuracy of approximately 90% and identifies variation in the outcome variable appropriately, as the ordering of average predictions given any true class corresponds to the ordering of the classes. Unfortunately, I did not use this model due to resource constraints (it would cost around 1,200 GPB to classify the 80,000 tweet-reply pairs in my dataset). Using this model would generate an outcome variable that is less noisy and more reliable than the outcome variables generated by the other two models. This translates into more precise and reliable regression coefficients, given the specification is correct.

Table 4: CONFUSION MATRIX LARGEST POLARIZATION CLASSIFIER

Actual \ Predicted	Polarized	Neutral	Anti-polarized
Polarized	8	10	0
Neutral	1	209	0
Anti-polarized	1	14	7

Notes: This table presents the raw Twitter interaction polarization classification accuracy of OpenAI's LLM DaVinci, fine-tuned to a training dataset of polarization-labeled Twitter interactions. Each cell shows the total number of validation records belonging to the given categories. The red numbers represent incorrect predictions, the black numbers represent correct predictions.

5.2. Classifying Interaction Topic

For the task of classifying the topic category for each tweet-reply pair and tweet without replies, I use the same text-completion model as for the task of classifying polarization. The fine-tune for this

[22]Which is Davinci at the time of writing, May 2023.

task was trained on around a thousand manually labelled tweet-reply pairs or tweets without replies and validated on around 250 such records. The accuracy of this model was around 80 percent, which is more promising here than it would in the previous classification task since these classes are more balanced. Below is an overview of the model's predictions, again in a confusion matrix, which allows one to see the decomposition of predictions given a true class.

Table 5: CONFUSION MATRIX TOPIC CLASSIFIER

Actual \ Predicted	Politics	Entertainment	Advertisement
Politics	49	12	2
Entertainment	5	203	7
Advertisement	1	16	24

Notes: This table presents the raw Twitter interaction topic classification accuracy of OpenAI's LLM Ada, fine-tuned to a training dataset of topic-labeled Twitter interactions. Each cell shows the total number of validation records belonging to the given categories. The red numbers represent incorrect predictions, the black numbers represent correct predictions.

As can be seen, this model features little misclassification. It performs better than the polarization classification model, perhaps because these categories are less imbalanced and are, in general, easier to identify in tweets, with certain words giving away the class rather than complex argumentative structures.

The following is an overview of the share of political tweets for each country and year. The share of political tweets rises for all countries from 2019 to 2020, perhaps related to COVID restrictions. Additionally, the share of tweets about politics in the US is higher than that in all other countries every year.

Table 6: THE SHARE OF POLITICAL TWEETS PER COUNTRY OVER TIME

Year	AUS	CAN	UK	US	Average
2018	0.11	0.12	0.12	0.15	0.12
2019	0.09	0.17	0.12	0.18	0.14
2020	0.12	0.24	0.15	0.34	0.21
2021	0.11	0.16	0.22	0.44	0.23
2022	0.11	0.15	0.09	0.22	0.14

Notes: This table presents the share of the collected tweets classified as topically political. I classify the collected tweets using OpenAI's Ada LLM that I fine-tune to a training dataset of topic-labeled tweets.

5.3. OLS

Before considering the results of an Ordinary Least Squares regression, I emphasize that the problematic nature of the polarization classification model causes variation in the outcome variable not to be identified, probabilistically, as shown in section 5.1. However, variation may be correctly identified to

some extent, with a strictly positive probability, as poor accuracy on the validation sample does not guarantee poor accuracy on the entire dataset[23]. Therefore, it may still be worth looking at the results from running OLS on the data.

Interpreting the coefficients of the regressions should be done with WhatsApp contamination to the Meta service coefficients in mind. Namely, Meta's Facebook and Instagram partially run on the same computer networks as WhatsApp (which is also part of Meta). This on its own means that the estimated effect comes closer to the true effect of interest than in cases of isolated outages, as there are fewer social media services for users to switch to during the outage. However, it also means that there can sometimes be simultaneous outages of Facebook, Instagram, and WhatsApp[24]. As WhatsApp is not a social media service in the classical sense, the treatment effect is contaminated by the effect of a reduction in the daily online communication between people. Controlling for WhatsApp outages would then be a way to deal with this concern, but due to high multicollinearity of those downtime variables, this would make coefficient estimates imprecise while only mildly improving the interpretation of the coefficients. Namely, one may argue the WhatsApp substitute is much more similar to the classic social media platforms than the average substitute consumed during a social media outage.

The table below shows the results for the regression of political polarization on the effective downtime per service without or with country and time fixed effects.

[23] Some may argue that poor accuracy on a validation set of 250 records is plenty of evidence to conclude that the classification model is poor. However, natural language on Twitter is extremely noisy as the breadth of possible types of conversation would imply.

[24] For example, during Meta's major 2021 outage during which Facebook, Instagram, and WhatsApp went down simultaneously for approximately six hours (The Verge 2021).

Table 7: EFFECTS OF PLATFORM DOWNTIME ON POLITICAL POLARIZATION

	Political polarization	
Estimated experienced downtime (hours)	(1)	(2)
Facebook	-0.0120	-0.00386
	(0.0330)	(0.0326)
Instagram	0.0292	0.0130
	(0.0334)	(0.0331)
YouTube	0.00914	-0.00785
	(0.0146)	(0.0147)
TikTok	0.00397	0.0150
	(0.0143)	(0.0137)
Snapchat	0.0182***	0.0139***
	(0.00345)	(0.00309)
Change in popularity Twitter	0.0177	0.0190
	(0.0164)	(0.0160)
Fixed effects		✓
Observations	6,530	6,530

Notes: This table presents the results of two regressions of political polarization as determined by the fine-tuned LLM model on the effective social media downtime. The observations are at the country × date × platform level. The fixed effects denote country and year fixed effects. I use heteroskedasticity and autocorrelation-consistent standard errors. The standard errors are provided in the parentheses, the significance symbols are * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$.

Regression (1) in table 7 shows that no coefficients of interest are of significance at the 5% level or lower, except for Snapchat. The positive coefficient says that the average marginal effect of an hour of Snapchat downtime, conditional on all other variables, increases polarization. This is equivalent to an additional hour of Snapchat being available reducing polarization relative to the average substitute.

The high significance on the Snapchat downtime variable can be explained by co-occurring outages, too, in this case ones that happen because of third-party server dependencies. Namely, Snapchat utilizes Google Cloud services (Yahoo 2021). Snapchat outages have been caused by Google Cloud issues which also affected many other online services, like Discord, Spotify, and numerous online games (The Verge 2021). Due to such instances, the estimated treatment effect is biased in the direction of the effect that a collective outage of Snapchat and these other online services have on the measure of polarization relative to the average substitute. Out of the services analyzed in this study, Snapchat seems to be the only one that had big contaminated outages multiple times. It is hard to separate the effect of shocks to usage of Snapchat from shocks to usage of other online services that are not social media platforms because the greatest outages of these services occurred in tandem.

A potential reason for a positive bias in Snapchat's coefficient is that during outages, which sometimes co-occur with Spotify outages, a large group of people is unable to consume as much music as they would like to, with consequential negative effects on people's moods and subsequently

increased political polarization as found on Twitter.

Regression (2) from table 6 shows that similar results hold for an OLS regression with both time and country fixed effects.

The coefficient on the Snapchat downtime variable means that the aggregate[25] effect of an additional hour of downtime of Snapchat increases polarization by close to one percent of the maximum range of the outcome variable[26]. Under the awkward assumption that the effect is linear up to unusually long outages, it would take about two full days of a Snapchat outage to bring a neutrally-polarized country to a fully polarized country, as estimated by this model.

Using table 8 displaying percentiles of the outcome variable, another interpretation of the magnitude of the regressor on Snapchat is that an hour more or less of an effective outage takes the median level (i.e., the 50th percentile) of political polarization either around 47 percentiles up or 35 percentiles down, respectively. This perhaps surprising finding is related to the overpowering quantity of tweets without a reply from a user, which are classified as neutral in their interaction and thus given a polarization score of `0`. Table 9 illustrates that the share of neutral tweets is relatively large for any country-year combination.

Table 8: DISTRIBUTION OF POLITICAL POLARIZATION

Percentile	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	97	98	99	100
Political polarization	-1.00	-0.80	-0.55	-0.40	-0.31	-0.24	-0.18	-0.14	-0.10	-0.07	-0.05	-0.04	-0.02	-0.01	0.00	0.00	0.01	0.04	0.12

Notes: This table represents the empirical distribution of political polarization to contextualize the regression coefficients.

Table 9: THE DECOMPOSITION OF COLLECTED TWEETS PER COUNTRY OVER TIME

	Single					Pairs				
Year	AUS	CAN	UK	US	Total	AUS	CAN	UK	US	Total
2018	20,344	77,279	37,688	179,647	314,958	1,950	6,019	4,412	8,302	20,683
2019	14,184	19,325	30,296	116,799	180,604	1,103	3,164	5,203	8,845	18,315
2020	3,842	7,244	13,345	70,411	94,842	417	851	2,485	7,456	11,209
2021	3,101	4,320	11,821	39,704	58,946	536	563	1,644	3,090	5,833
2022	3,292	5,805	10,349	65,248	84,694	577	1,247	3,943	4,470	10,237
Total	45,742	114,905	105,239	477,748	743,634	4,831	12,105	18,355	32,861	68,152

Notes: This table represents the decomposition of the collected tweets. I use all these collected tweets in all analyzes involving tweets in this paper.

I should also note that Twitter users delete around 11% of tweets (Bhattacharya et al. 2022). These

[25] An increase of an outage by one hour increases the estimated experienced downtime equivalent by one hour over that date and the four dates after, as per equation (5) on page 10. Therefore, by linearity, the aggregate effect of an additional hour of downtime is equivalent to the effect of one additional hour of the experienced downtime equivalent.

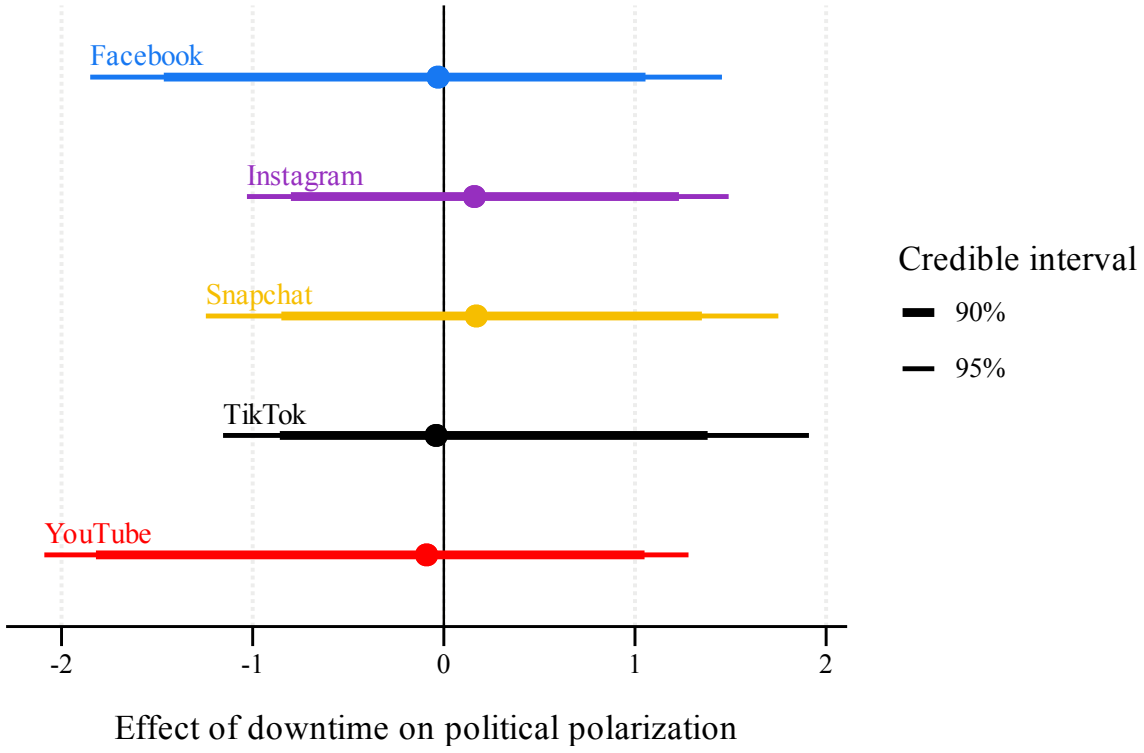
[26] Namely, as discussed in the "Quantifying Polarization" subsection, if all interactions on a day are polarized, the polarization score is `1` and if all interactions on a day are anti-polarized, the polarization score is `-1`

tweets can unfortunately not be recovered. This means that some of these tweets without replies may have had a reply at first which was polarized because deleted tweets are often part of heated discussions (Tweet Deleter 2021). This means that the outcome variable may not display as much variation as there is in truth, potentially causing attenuation bias of the regressor coefficients. However, it can also be that most of the deleted tweets are those posted by bot accounts that violate Twitter's terms and conditions.

5.4. Bayesian Regression

The below is a figure displaying the 90 and 95% highest posterior density regions for the parameters of interest as estimated by the Bayesian regression.

Figure 2: Posterior Distribution of Platform Coefficients



Notes: This figure displays the posterior distribution of the effect of social media downtime on political polarization by platform-specific coefficient used in the hierarchical model of political polarization. The dots represent posterior means. I provide an explanation of the model specification in section 4.2.

The results differ from the OLS results in the sense that all coefficients are insignificant, and to a greater extent. The greater confidence region of the parameters can be explained by the fact that the variance of the posterior distribution of the parameters is not just determined by the variance on the error term but also by the variance terms of the other parameters, which partially covary.

The MCMC algorithm does not converge, implying the data is too sparse to estimate relative to the number of parameters and the variance allowed on the parameters. Given that the modelling decisions are arguably modest but intuitively well-supported, this may say something about the validity of the

OLS specification results. Namely, given that the variation of the outcome variable is generated by a balanced confusion of the classification model between the non-neutral classes, which numerically absolutely differ from the neutral class by the same amount, 1, the outcome variable should not actually be that much correlated to anything. This would make the significance of some of the OLS regressors dubious. OLS's regressors not covarying by assumption and being non-regularized versus Bayesian's option to have them covary and be regularized can explain why one doesn't display the expected result whereas the other does. Still, it may be that anti-polarized interactions are given more weight when being aggregated, meaning that in periods of increased proportions of non-neutral interactions, polarization levels may be lower. This in turn allows correlations to be identified between the outcome variable of political polarization and the regressors.

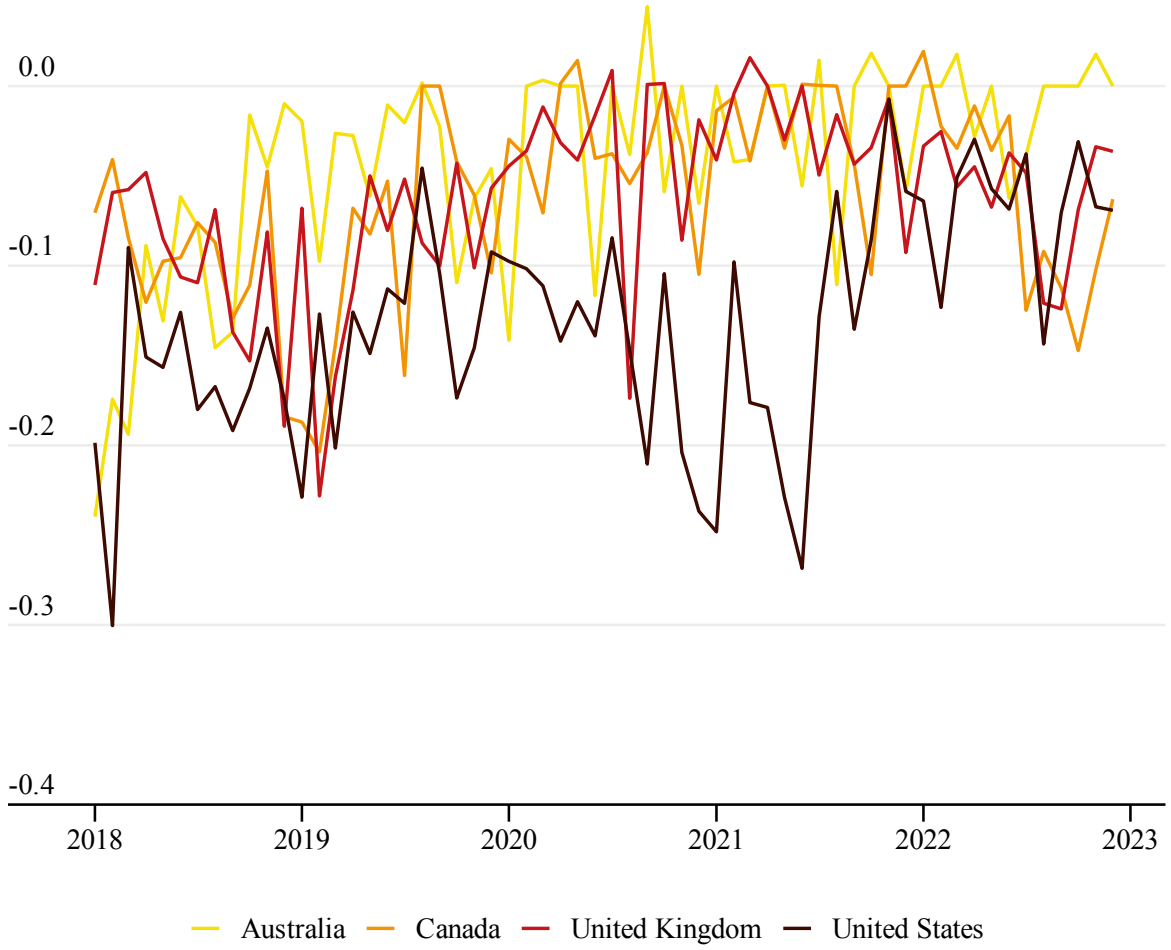
5.5. Robustness Checks

5.5.1. Time Confounding

The first robustness check is testing whether there exists a correlation between the outage severity of the social media platforms analyzed and time, and whether average polarization is correlated with time. This will motivate the inclusion or exclusion of time-fixed effects.

Below is the behaviour of polarization in political interactions as defined in this study per country over time.

Figure 3: POLITICAL POLARIZATION PER COUNTRY OVER TIME



Notes: This figure presents the monthly means of estimated political polarization, by country. I calculate polarization by day by aggregating the polarization scores of Twitter interactions that are labeled to be topically political. I describe the weights used in this aggregation excercise in detail in section 3.3.

There seems to be an overall upward time trend in political polarization, as classified by the model, for all countries except perhaps the US, which seems to have a decreased level of political polarization coinciding with a larger share of tweets being categorized as political around when COVID-19 restrictions were most severe. This means that a similar linear trend in the number of outages over time can be a source of bias because the COVID pandemic also took place during the later years of this dataset (the third and fourth out of all five years), which may have impacted polarization.

Below is a covariance matrix between the average outage severity of each service with the number of elapsed years that have elapsed since 2018.

Table 10: COVARIANCE YEARS ELAPSED AND OUTAGE SEVERITY

Platform	Covariance
Facebook	0.01
Instagram	0.01
TikTok	-0.02
YouTube	0.01
Snapchat	0.02

Notes: This table presents the covariance between the number of years elapsed since the start of the panel with the average severity of outages in that year, by platform.

It can be seen that the magnitude of the largest covariance is 0.02, meaning that an average increase in the years elapsed by one is expected to coincide with an average change in the effective downtime by 0.02, which corresponds to 1 minute and 12 seconds. Over five years, this would only translate to 6 minutes of a difference in the cumulative downtime for the service in question. This is insignificant in comparison to the cumulative downtime of services, which ranges from 2.5 hours (TikTok) to 7.5 hours (Snapchat). Of course, this is only the linear trend that can be identified, so this table does not disprove the necessity of time fixed effects, but may alleviate the mentioned concern that exists if one expects polarization to contain a non-negative linear time trend caused by something other than social media outages.

5.5.2. Robustness of Downtime Transformation

Next I display the OLS results with the estimated equivalent outage duration variable and its lagged versions, to see whether the results of the regression with the effective experienced downtime variables are different because of the transformed variable created from the multiple corresponding lagged variables.[27] As is apparent, the results here are comparable as they also feature similar insignificance of coefficients and significance on the Snapchat downtime variable. The significance of the Snapchat coefficient increases the longer the lag is. This may imply that the effect of a Snapchat outage (and that of the co-occurring outages of other online services as I discussed before) takes time to have a tangible effect on polarization, supporting the view that shocks in polarization spread over a society through diffusing processes of interactions over a few days. I should also note that I included the fifth and sixth variable downtime variables but cut these from the regression table. This part of the table featured significance in the coefficient on Snapchat downtime for the 5th lag and insignificant coefficients otherwise, providing similar insight.

[27]For the results of other OLS specifications, including a regression of polarization as measured in entertainment interactions on Twitter and a regression of political polarization as classified by the lower-accuracy embeddings model, see the appendix sections 8.5 and 8.6.

Table 11: EFFECTS OF PLATFORM DOWNTIME ON POLITICAL POLARIZATION

Estimated equivalent downtime	Political polarization (1)	(2)
Facebook	0.00806	0.0186
	(0.0746)	(0.0762)
Instagram	-0.0554	-0.0715
	(0.0532)	(0.0521)
YouTube	-0.0637	-0.0766*
	(0.0365)	(0.0366)
TikTok	-0.0466	-0.0397
	(0.0335)	(0.0331)
Snapchat	0.00633	0.00247
	(0.0145)	(0.0142)
Facebook L1	-0.00913	-0.00693
	(0.0406)	(0.0408)
Instagram L1	0.0289	0.0204
	(0.0405)	(0.0409)
YouTube L1	0.00500	-0.00806
	(0.0194)	(0.0182)
TikTok L1	0.00475	0.00960
	(0.0253)	(0.0242)
Snapchat L1	0.0149*	0.0125*
	(0.00623)	(0.00554)
Facebook L2	-0.0122	-0.00930
	(0.0542)	(0.0539)
Instagram L2	0.0236	0.0148
	(0.0562)	(0.0556)
YouTube L2	0.0339	0.0200
	(0.0215)	(0.0216)
TikTok L2	0.0221	0.0275
	(0.0340)	(0.0316)
Snapchat L2	0.0182***	0.0136**
	(0.00453)	(0.00449)
Facebook L3	0.00966	0.0167
	(0.0444)	(0.0447)
Instagram L3	0.00511	-0.00701
	(0.0438)	(0.0435)
YouTube L3	-0.0423	-0.0546
	(0.0317)	(0.0296)
TikTok L3	-0.0528	-0.0446
	(0.0335)	(0.0328)
Snapchat L3	0.00121	0.000861
	(0.0104)	(0.0102)
Facebook L4	-0.0578	-0.0462
	(0.0357)	(0.0354)
Instagram L4	0.0711*	0.0540
	(0.0353)	(0.0349)
YouTube L4	0.0123	0.000327
	(0.0125)	(0.0103)
TikTok L4	0.0584*	0.0653*
	(0.0265)	(0.0277)
Snapchat L4	0.0197**	0.0155*
	(0.00730)	(0.00669)
Fixed effects		✓
5th and 6th variable lags	✓	✓
Change in Twitter usage control	✓	✓
Observations	6,530	6,530

Notes: This table presents the results of two regressions of political polarization as determined by the fine-tuned LLM model on the effective social media downtime. The observations are at the country × date × platform level. The fixed effects denote country and year fixed effects. I use heteroskedasticity and autocorrelation-consistent standard errors. The standard errors are provided in the parentheses, the significance symbols are * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$.

6. Discussion

6.1. Reducing Volatility

As I emphasize, the main problem I encountered in this research is the inaccuracy of the classification of polarization in tweet-reply pairs, leading to the outcome variable being uninformative with a high probability. The OpenAI model that I used for this classification task is not the best model meant for this purpose that OpenAI offers. Namely, the quality of text completion seems to be (economically significantly) linearly correlated to model size (Ray 2020), and the size of the model used, GPT-3 Ada, is about four times as small as GPT-3's Davinci model. This means that predictions with the GPT-3 Davinci model would have likely been better at classifying and would have maybe allowed for the correct identification of variation in the outcome variable of polarization, as the estimations of the model on the testing data suggest. However, I did not use this latter version of the model due to the higher associated cost of classification. Namely, the classification of the around 80k tweet-reply pairs cost around 20 GBP with GPT-3 Ada, but would have cost around 1,600 GBP with GPT-3 Davinci.

Additionally, even with more accurate classification of polarization, there is a concern about the outcome variable being too volatile to give the coefficient estimates of interest the standard errors that allow one to make claims about any causal relations with significant confidence. Namely, filtering the non-political interactions out of the dataset of tweet-reply pairs, only about 20,000 tweet-reply pairs remain. In 1,825 days, on average there are approximately eleven tweet-reply pairs per day with which you can identify variation in polarization. Of course, there are about 800,000 tweets in the dataset with no replies out of which approximately 250,000 are classified to be about politics, but these are all labelled as `neutral`, so they do not really contribute to variation in the outcome variable. A way to increase power and reduce this worry is therefore to increase the number of tweets analyzed per day. However, after filtering on country and the time range 2018-2022, I did, in fact, scrape all available tweets on Twitter before adding them to the dataset. Nevertheless, as I discussed in the tweet collection section, it is this geo filter that removes most tweets per country because only a small fraction of tweets are geocoded. Removing this filter would increase the sample greatly. Given that you are likely to obtain the best polarization classification results on English text (Muennighoff et al. 2022), and Twitter labels tweets' language, and you can filter on Twitter's labels when scraping, taking this approach may take care of this concern. Outages will then have to be defined based on all English-speaking countries, weighting by population and Twitter popularity among that population in a sophisticated manner.

6.2. Search Engine Bias

A potential source of bias with using Google Trends is that not all people browse on Google. Bing users may use YouTube at different times during the day than Google users, and therefore browsing activity related to a YouTube outage concentrated during a period in which Bing users consume much YouTube but Google users consume little will be missed by the Google Trends data. This concern

is unlikely to be a major one, though, as 90-93% of people use Google over other search engines, like Bing, in the English-speaking countries of interest (Statcounter 2023). Since the usage of search engines other than Google is small, it would take a substantial difference in usage patterns of the users of different search engines to generate a bias of significant magnitude.

6.3. Methodological Limitations

I should note that the methodology of this paper measures a short-term effect. Widespread social media outages have never lasted more than a day. Perhaps it takes longer than a couple of hours of social media disuse for an effect to occur that is significant enough to be measured systematically. It is not clear what a straightforward approach would be to quantify the causal effect of a sustained drop in social media usage. Long-lasting RCTs would underestimate the effect due to attenuation bias caused by contamination through interactions between the non-treated and treated populations, while event studies suffer from confounding or the local nature of their effect estimates whenever stripped from endogeneity.

7. Conclusion

In this study, I proposed the usage of a new measure of polarization that quantifies the divisiveness in the average interaction in society by looking at the divisiveness of the average Twitter interaction in that society. Its computation does not rely on predefining groups of affiliation and is feasible by leveraging large pre-trained language models of text continuation, like OpenAI's GPT-3. I then estimate the causal effect of social media usage on political polarization to be insignificant, using Google Trends data on the popularity of search queries like *"is youtube down"* to estimate the severity of outages over time. Nevertheless, the usage of bigger, more expensive classification models of OpenAI identify variation in the outcome variable more accurately and give rise to more precise and reliable coefficient estimates. Lastly, I find that a neural network polarization classifier trained on a large dataset of labelled vector-embedded Reddit conversations from a selection of forums is not accurate in predicting polarization from vector-embedded Twitter conversations.

References

1.5M clients, 1B deleted tweets statistics — TweetDeleter — tweetdeleter.com (n.d.). https://tweetdeleter.com/blog/1-5m-clients-1b-deleted-tweets-tweet-deleter-reveals-statistics/. [Accessed 03-May-2023].

About — Sensor Tower — sensortower.com (n.d.). https://sensortower.com/about. [Accessed 26-Apr-2023].

About our approach to recommendations — help.twitter.com (n.d.). https://help.twitter.com/en/rules-and-policies/recommendations. [Accessed 28-Apr-2023].

Advanced filtering for geo data — developer.twitter.com (n.d.). https://developer.twitter.com/en/docs/tutorials/advanced-filtering-for-geo-data. [Accessed 22-Apr-2023].

Allcott, Hunt and Matthew Gentzkow (2017). "Social media and fake news in the 2016 election". In: *Journal of economic perspectives* 31.2, pp. 211–236.

Barberá, Pablo (2014). "How social media reduces mass political polarization. Evidence from Germany, Spain, and the US". In: *Job Market Paper, New York University* 46, pp. 1–46.

Bhattacharya, Parantapa, Saptarshi Ghosh, and Niloy Ganguly (2022). *Analyzing Regrettable Communications on Twitter: Characterizing Deleted Tweets and Their Authors*. arXiv: 2212.12594 [cs.SI].

Brown, Tom B, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (May 2020). "Language Models are Few-Shot Learners". In: arXiv: 2005.14165 [cs.CL].

Brownlee, Jason (n.d.). *How Much Training Data is Required for Machine Learning*. https://machinelearningmastery.com/much-training-data-required-machine-learning/. [Accessed 26-Apr-2023].

Countries with most Twitter users 2022 — Statista — statista.com (n.d.). https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/. [Accessed 20-Apr-2023].

Demszky, Dorottya, Nikhil Garg, Rob Voigt, James Zou, Matthew Gentzkow, Jesse Shapiro, and Dan Jurafsky (2019a). *Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings*. arXiv: 1904.01596 [cs.CL].

— (2019b). "Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings". In: *CoRR* abs/1904.01596. arXiv: 1904.01596. URL: http://arxiv.org/abs/1904.01596.

Dinesh, Shradha and Meltem Odabaş (n.d.). *8 facts about Americans and Twitter as it rebrands to X — pewresearch.org*. https://www.pewresearch.org/short-reads/2023/07/26/8-facts-about-americans-and-twitter-as-it-rebrands-to-x/. [Accessed 29-11-2023].

Estimating Twitter's Bot-Free Monetizable Daily Active Users (mDAU) — Similarweb Research — similarweb.com (n.d.). https://www.similarweb.com/amp/blog/insights/social-media-news/twitter-bot-research/. [Accessed 24-Apr-2023].

Facebook, Instagram and other Meta services suffer massive global outage, services being restored slowly — firstpost.com (n.d.). https://www.firstpost.com/world/facebook-instagram-and-other-meta-services-suffer-massive-global-outage-services-being-restored-slowly-12050212.html. [Accessed 22-Apr-2023].

Falkenberg, Max, Alessandro Galeazzi, Maddalena Torricelli, Niccolò Di Marco, Francesca Larosa, Madalina Sas, Amin Mekacher, Warren Pearce, Fabiana Zollo, Walter Quattrociocchi, and Andrea Baronchelli (2022). "Growing polarization around climate change on social media". In: *Nature Climate Change* 12.12, pp. 1114–1121. DOI: 10.1038/s41558-022-01527-x. URL: https://doi.org/10.1038/s41558-022-01527-x.

FAQ about Google Trends data - Trends Help — support.google.com (n.d.). https://support.google.com/trends/answer/4365533?hl=en. [Accessed 24-Apr-2023].

Garimella, Kiran, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis (July 2015). "Quantifying controversy in social media". In: arXiv: 1507.05224 [cs.SI].

Gelman, Andrew (n.d.). *Prior Choice Recommendations — github.com*. https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations. [Accessed 26-Apr-2023].

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy (2019). "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech". In: *Econometrica* 87.4, pp. 1307–1340. DOI: https://doi.org/10.3982/ECTA16566. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA16566. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA16566.

Google Cloud leads major outage as Snapchat and Spotify down — uk.finance.yahoo.com (n.d.). https://uk.finance.yahoo.com/news/google-cloud-leads-major-outage-195701204.html. [Accessed 26-Apr-2023].

Google services outages - Wikipedia — en.wikipedia.org (n.d.). https://en.wikipedia.org/wiki/Google_services_outages. [Accessed 24-Apr-2023].

Heath, Alex (n.d.). *Facebook is scrambling to fix massive outage — theverge.com*. https://www.theverge.com/2021/10/4/22709575/facebook-outage-instagram-whatsapp. [Accessed 26-Apr-2023].

Hetherington, Marc J and Thomas J Rudolph (2020). *Why Washington won't work: Polarization, political trust, and the governing crisis*. University of Chicago Press.

Infographic: 80% Of Twitter's Users Are Mobile — statista.com (n.d.). https://www.statista.com/chart/1520/number-of-monthly-active-twitter-users/#:~:text=Twitter%

20presented%20its%20latest%20results,Twitter%20via%20their%20mobile%20device.. [Accessed 26-Apr-2023].

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J. Westwood (2019). "The Origins and Consequences of Affective Polarization in the United States". In: *Annual Review of Political Science* 22.Volume 22, 2019, pp. 129–146. ISSN: 1545-1577. DOI: https://doi.org/10.1146/annurev-polisci-051117-073034. URL: https://www.annualreviews.org/content/journals/10.1146/annurev-polisci-051117-073034.

Kubin, Emily and Christian von Sikorski (2021). "The role of (social) media in political polarization: a systematic review". In: *Annals of the International Communication Association* 45.3, pp. 188–206. DOI: 10.1080/23808985.2021.1976070. eprint: https://doi.org/10.1080/23808985.2021.1976070. URL: https://doi.org/10.1080/23808985.2021.1976070.

Lee, Changjun, Jieun Shin, and Ahreum Hong (2018). "Does social media use really make people politically polarized? Direct and indirect effects of social media use on political polarization in South Korea". In: *Telematics and Informatics* 35.1, pp. 245–254. ISSN: 0736-5853. DOI: https://doi.org/10.1016/j.tele.2017.11.005. URL: https://www.sciencedirect.com/science/article/pii/S0736585317305208.

Levy, Ro'ee (Mar. 2021). "Social Media, News Consumption, and Polarization: Evidence from a Field Experiment". In: *American Economic Review* 111.3, pp. 831–70. DOI: 10.1257/aer.20191777. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20191777.

McCoy, Jennifer, Tahmina Rahman, and Murat Somer (2018). "Polarization and the Global Crisis of Democracy: Common Patterns, Dynamics, and Pernicious Consequences for Democratic Polities". In: *American Behavioral Scientist* 62.1, pp. 16–42. DOI: 10.1177/0002764218759576. eprint: https://doi.org/10.1177/0002764218759576. URL: https://doi.org/10.1177/0002764218759576.

Muennighoff, Niklas, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel (Nov. 2022). "Crosslingual Generalization through Multitask Finetuning". In: arXiv: 2211.01786 [cs.CL].

Neelakantan, Arvind, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng (Jan. 2022). "Text and Code Embeddings by Contrastive Pre-Training". In: arXiv: 2201.10005 [cs.CL].

New and improved embedding model — openai.com (n.d.). https://openai.com/blog/new-and-improved-embedding-model. [Accessed 26-Apr-2023].

OpenAI API — platform.openai.com (n.d.[a]). https://platform.openai.com/docs/guides/fine-tuning. [Accessed 26-Apr-2023].

OpenAI API — platform.openai.com (n.d.[b]). https://platform.openai.com/docs/guides/completion/prompt-design. [Accessed 22-Apr-2023].

OpenAI's gigantic GPT-3 hints at the limits of language models for AI — zdnet.com (n.d.). https://www.zdnet.com/article/openais-gigantic-gpt-3-hints-at-the-limits-of-language-models-for-ai/. [Accessed 26-Apr-2023].

Pougué-Biyong, John, Valentina Semenova, Alexandre Matton, Rachel Han, Aerin Kim, Renaud Lambiotte, and Doyne Farmer (2021). "DEBAGREEMENT: A comment-reply dataset for (dis)agreement detection in online debates". In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. URL: https://openreview.net/forum?id=udVUN__gFO.

Search Engine Market Share Worldwide — Statcounter Global Stats — gs.statcounter.com (n.d.). https://gs.statcounter.com/search-engine-market-share. [Accessed 24-Apr-2023].

Short, Cole E. and Jeremy C. Short (2023). "The artificially intelligent entrepreneur: ChatGPT, prompt engineering, and entrepreneurial rhetoric creation". In: *Journal of Business Venturing Insights* 19, e00388. ISSN: 2352-6734. DOI: https://doi.org/10.1016/j.jbvi.2023.e00388. URL: https://www.sciencedirect.com/science/article/pii/S2352673423000173.

Status overview — downdetector.co.uk (n.d.). https://downdetector.co.uk/about-us/. [Accessed 24-Apr-2023].

Technology Primer: Social Media Recommendation Algorithms — belfercenter.org (n.d.). https://www.belfercenter.org/publication/technology-primer-social-media-recommendation-algorithms. [Accessed 22-Apr-2023].

Twitter Users, Stats, Data, Trends, and More — DataReportal – Global Digital Insights — datareportal.com (2023). [Accessed 29-11-2023].

Wike, Richard, Laura Silver, Janell Fetterolf, Christine Huang, Sarah Austin, Laura Clancy, and Sneha Gubbala (2022). *Social Media Seen as Mostly Good for Democracy Across Many Nations, But U.S. is a Major Outlier — pewresearch.org*. https://www.pewresearch.org/global/2022/12/06/social-media-seen-as-mostly-good-for-democracy-across-many-nations-but-u-s-is-a-major-outlier/. [Accessed 27-11-2022].

8. Appendix

8.1. Additional Information Tweet Collection

The way I select a reply to a tweet in section 3.1 is randomly picking one out of all replies to the tweet that is not from the user that posted the tweet, meaning that the interaction type of the replies in these tweet-reply pairs corresponds on average to the average interaction type of all replies to the respective tweets. I do not include quotes[28] in the dataset in order to avoid language model classifier issues induced by missing context for the tweet whenever a quote was a quote itself, and so forth. I leave out replies to replies because for some replies there could be too little context for the language model classifier to determine the interaction type, and if I gave all preceding interactions from the conversation thread of any single reply to the language model classifier then costs related to model classification would have been too large.

I do not analyze interactions like those contained in pairs of replies to replies or quotes to tweets. Ideally, of course, you would label the interaction type for every possible Twitter interaction and aggregate all of them in some way. However, the labelling task and aggregating process become complex quickly as Twitter allows for many different types of interaction threads, while it is also hard to determine when one conversation topic evolves into another[29].

Lastly, contrary to the polarization aggregation method based on likes in this paper, an aggregation method based on taking retweets as endorsement is also possible. However, replies to tweets rarely tend to be retweeted[30], meaning this approach may give overly noisy results.

8.2. Further Motivation Choice Classification Model

There are numerous language models available that can classify the interaction type contained in complex verbal interactions for the tweet-reply pairs in the data. There exist probabilistic methods that have been used in the literature, including Gentzkow's (2019) which has been used in Ro'ee (2019), for example. This method in particular works only when affiliation to any of two groups is predetermined and looks at the extent of the difference in word usage between two groups. As I mention in the theoretical framework in section 2, the analysis in this study does not predetermine users' groups, and can therefore not be used. Moreover, one would probably like to label an interaction to be of a given polarization grade whenever we, as humans, would mostly agree that it is of that particular grade. Methods like Gentzkow's do not distinguish between the order of words appearing in a piece of text and will therefore not be suited to the ambition of identifying the interaction type between a tweet

[28] A quote is defined to be a pair of tweets consisting of a quoted tweet and the tweet quoting it.

[29] For example, every time a tweet is quoted by a user, that user shares the quoted tweet and their reply to that tweet with all their followers. Replies on this tweet can in turn follow, and one of these replies may in turn start a new thread of replies. It is not hard to see why networks created from Twitter conversation can grow to become large in size (Garimella et al. 2015).

[30] In the dataset of scraped tweets and replies, tweets with replies have 17.2 retweets on average while replies have 0.1 retweets on average.

and its reply like humans would, as the interaction type depends on word choice and order in both the tweet and the reply.

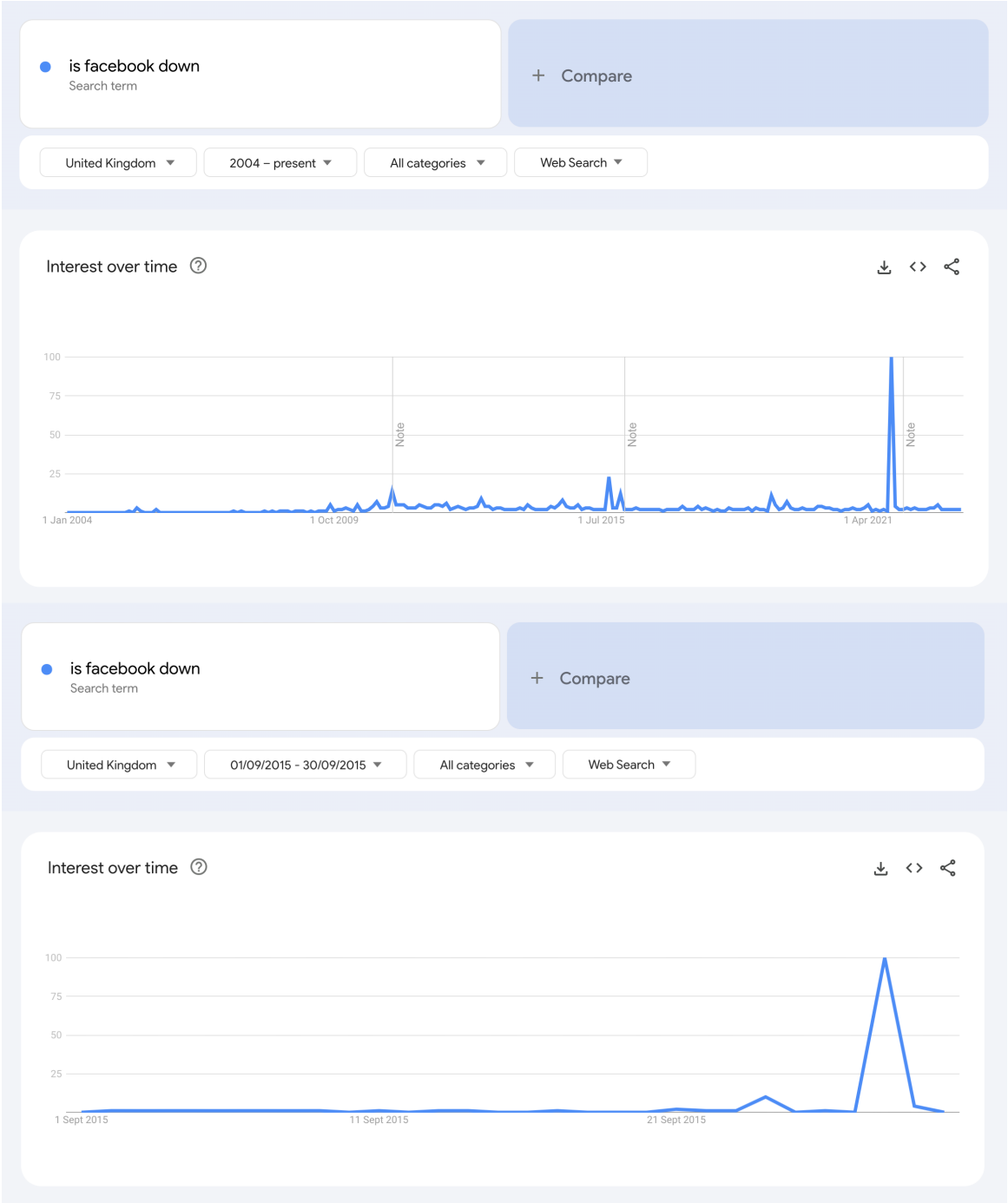
Lastly, the way in which I used the classification model is not the only one. There are many other ways to construct a prompt given the text contained in `[tweet]` and `[reply]` that satisfies the requirements as given in the text. I do not investigate the robustness of these relative to others due to time and resource constraints. Additionally, there is no consensus on how best to structure prompts for text-completion models (C. Short and J. Short 2023). Although prompt changes have a smaller effect the better language models become and adjusting prompts with contemporary models does not always improve results drastically, this can be looked at in further studies.

8.3. Technicalities Google Trends Data

One technical difficulty with scraping data from Google Trends is that daily data is only available per month, and is scaled to have maximum popularity at 100 for that same month. Data from two different months about search queries related to outages of the same service in the same country are therefore not comparable unless you scale by the relative popularity of that search query for each of those months, which you can do by asking Google Trends to give you data on the popularity of that search term over the history of the Google Trends database by months[31]. I next include an illustration of what Google Trends returns for each search (top image: all-time, bottom image: specific month). To get the popularity of a search in a given day, one will calculate the popularity of the query in a given day as a fraction of the sum of the popularity of that search query that month before scaling by the popularity of the specific search query of that month as a whole. In essence, you multiply the relative proportion of Google searches about a certain query of a month by the relative proportion of Google searches about that query of a specific day in that month, which gives you the relative proportion of Google searches about a certain query in a specific day.

[31]This is also the setting for which Google Trends does not offer daily data, otherwise this scaling approach was not needed to begin with: one would simply query the daily data over the desired time range with the filters of interest

Figure 4: GOOGLE TRENDS DATA COLLECTION OPTIONS



Notes: Upper picture: monthly relative popularity data of the search query "*is facebook down*" in the United Kingdom, from the start of Google's database (January 1st, 2004). Lower picture: daily relative popularity data of the same search query in the same country for the month of September in 2015. Notice that for the bottom picture, the maximum is 100, too, while the maximum popularity of the search query does not occur in a month in 2015, but in a month in 2021, as shown in the upper picture. This is because of the scaling that Google Trends applies to the data returned for each query.

8.4. Downtime Data Alternatives

There were many downtime data alternatives I considered before deciding to use Google Trends data. There are a couple of reasons why I chose Google Trends data that I think are important to point out.

First of all, using only the social media outages over the last few years that are covered widely in news articles online is an easy way to collect downtime data, but is not the best possible, as many smaller outages happen each year that are not always covered by the media or are hard to find, if so. Moreover, if these smaller outages are covered, not all media articles contain information on the duration of these outages or an estimation of the number of users affected. For these reasons, one may start to think about alternative ways to gather data on outages.

What happens when a service goes down is that users that can not access the service may query the web to confirm whether that service is in fact down. Sometimes, users report having problems with a service. `downdetector.com` shows up first after Google queries, usually, and it is the largest website in the world for user-induced outage detection (Downdetector 2023). Downtime data on `downdetector.com` is only available for the last 24 hours from their website. One could start scraping this data every day to compile a dataset, but one would not be able to look further back than 24 hours from the time they start this process. Luckily, Downdetector posts tweets whenever they detect a large enough number of users reporting a platform outage to conclude that the service is having at least a partial outage, for each of the major English-speaking languages and for all major social media services. These tweets are publicly available and go back to at least the beginning of 2018, the start of the period of interest for this study.

If one wants to use these tweets to estimate the duration and severity of outages, looking at likes, retweets, and replies may be a sensible step to take. Namely, one can expect people that have difficulties with using some social media services to browse whether the service in question is down for other people, too. After a search query like "is tiktok down", for example, they may press a few links and arrive at Downdetector's Twitter page, after which they may like or retweet the post reporting an issue with the service in question. One problem with this approach is that in different periods of time, Google and other search engines may be more or less likely to display links that lead to Downdetector's Twitter account. Additionally, using the number of likes on a service outage's tweet to determine the outage's severity is complicated due to users of other platforms like Twitter using Twitter to different degrees. Estimations of outage durations and severity will then feature more variation for the services with a larger following on Twitter, holding all other variables fixed. There are also issues of bias related to demographic differences in user bases of different social media platforms caused by the propensity to like a tweet reporting an outage corresponding to the one experienced by a given user of that platform. It is not easy to deal with the mentioned biases due to data on the confounding variables not being available.

Lastly, I could have aggregated Google Trends data of different search queries with the search query that I use in this study. Still, it is not clear which other search queries than the one I use (i) are short enough to be less volatile than longer, more uncommon search queries, and (ii) contain a similar questioning element to it that is desirable when trying to identify people that are experiencing issues with some social media service versus people that want to read the news about a social media outage.

8.5. Results Entertainment Category

The below table displays the effects of social media platform downtime on polarization of discussions related to entertainment. The purpose of this additional regression is to check whether the platform downtime effects on polarization are heterogeneous by discussion topic the polarization is measured from.

Table 12: EFFECTS OF PLATFORM DOWNTIME ON POLARIZATION IN ENTERTAINMENT

	Polarization in entertainment discussions	
Estimated experienced downtime (hours)	(1)	(2)
Facebook	0.00237	0.00594
	(0.0112)	(0.0112)
Instagram	-0.00101	-0.00511
	(0.0108)	(0.0109)
YouTube	-0.00412	-0.00684
	(0.00721)	(0.00721)
TikTok	0.00427*	0.00433*
	(0.00203)	(0.00203)
Snapchat	0.00288*	0.00285*
	(0.00128)	(0.00128)
Change in popularity Twitter	-0.00543	-0.00546
	(0.00544)	(0.00542)
Fixed effects		✓
Observations	7,386	7,386

Notes: This table presents the results of two regressions of polarization in entertainment-related discussions as determined by the fine-tuned LLM model on the effective social media downtime. The observations are at the country × date × platform level. The fixed effects denote country and year fixed effects. I use heteroskedasticity and autocorrelation-consistent standard errors. The standard errors are provided in the parentheses, the significance symbols are * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$.

Although two variables are now significant, in contrast to just one in table 7, the magnitude of these coefficients is economically insignificant, really. Namely, it would take over ten days of downtime to bring a country that is neutral in entertainment polarization to a country that is either fully polarized or fully anti-polarized in entertainment, relative to only two days that it takes for political polarization by table 7. These results may be interpreted in a way that polarization in entertainment-related interactions is not really influenced by social media usage, but again the issues related to the quantification of the outcome variable should be kept in mind.

8.6. Results Embeddings Classification

The following table the OLS results related to political polarization as classified by the embeddings model.

Table 13: EFFECTS OF PLATFORM DOWNTIME ON POLITICAL POLARIZATION

	Political polarization	
Estimated experienced downtime (hours)	(1)	(2)
Facebook	-0.0349	-0.0283
	(0.0592)	(0.0574)
Instagram	0.0656	0.0390
	(0.0575)	(0.0556)
YouTube	0.00649	-0.0236
	(0.0243)	(0.0225)
TikTok	-0.0543*	-0.0375
	(0.0255)	(0.0259)
Snapchat	0.0356***	0.0270**
	(0.00924)	(0.00946)
Change in popularity Twitter	-0.00132	0.00104
	(0.0284)	(0.0282)
Fixed Effects		✓
Observations	6,530	6,530

Notes: This table presents the results of two regressions of political polarization as determined by the embeddings model on the effective social media downtime. The observations are at the country × date × platform level. The fixed effects denote country and year fixed effects. I use heteroskedasticity and autocorrelation-consistent standard errors. The standard errors are provided in the parentheses, the significance symbols are * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$.

The coefficients are much larger in magnitude than for the regression of political polarization as classified by the text-completion model in table 7. This can be explained by the fact that variation in the outcome variable is actually identified correctly by the embeddings model, despite its low accuracy. The standard errors are relatively higher because of this low accuracy, but the outcome is still determined. The interpretation of the coefficient on TikTok is that an additional hour of the platform being down aggregately reduces polarization by 2.5% of its maximum range in table, which is determined by the extremes of political polarization levels of either `1` and `-1`. Another interpretation of the coefficient is that a change in one hour of TikTok downtime would bring a median level of political polarization 39 percentiles down or 49 percentiles up, see table 8. The result for TikTok is not robust to time and country fixed effects, though. The result for Snapchat is significant and similar to the results in table 7, though the magnitude of the coefficient is about 50% larger.